

## A normális eloszlás paramétereire vonatkozó próbák; illeszkedésvizsgálat

Az alábbi próbák akkor használhatók, ha

- a megfigyelések függetlenek, és feltételezhetjük, hogy normális eloszlásúak vagy
- a megfigyelések függetlenek, véges szórású eloszlásból származnak, és a minta mérete, azaz  $n$  "elég nagy", például  $n \geq 100$ ; ez a **centrális határeloszlástételen** múlik: tetszőleges véges szórású, független azonos eloszlású valószínűségi változók átlagának eloszlása normális eloszláshoz hasonló, ha nagy a mintaelemszám
- ugyanakkor, **túl nagy mintaelemszám** esetén a próba túlságosan érzékennyé válik, kis eltérést is szignifikánsnak jelez – ezért fontos az erőfüggvény, abból vehetjük észre, hogy túl nagy a mintaelemszám, hogy az erőfüggvény túl kicsi
- **$z$ -próba** (vagy  $u$ -próba): **várható értékre** vonatkozó hipotézis esetén, ha **a  $\sigma$  szórás ismert** – egymintás esetben legerősebb próba
- **$t$ -próba** (vagy Student-próba): **várható értékre** vonatkozó hipotézis esetén, ha **a  $\sigma$  szórás nem ismert** (csak az  $s_n^*$  tapasztalati szórás)
- **$F$ -próba**: **szórásra** vonatkozó hipotézis esetén

Már láttuk a  $z$ -próbákat, illetve egymintás esetben a  $t$ -próbát. Nézzük, mi történik, ha két mintát szeretnénk összehasonlítani, azonban nem ismertek a szórások.

### 1. Kétmintás, egyoldali, párosítatlan Student-féle $t$ -próba

Ez a próba a **várható érték összehasonlítására** szolgál **azonos szórás esetén** (two-sample one-sided unpaired Student  $t$ -test).

Két különböző csoportra jellemző várható értékek összehasonlításánál is gyakori, hogy a szórásokat valójában nem ismerjük, csak a mintából kiszámítható korrigált tapasztalati szórásokat. Viszont Student-féle  $t$ -próba feltételei közé tartozik az is, hogy a szórások, bár nem ismertek, a két csoportban azonosak. Például: két tárgy, élőlény stb. valamilyen jellemzőjét mérjük ugyanazzal a mérési eljárással, és feltesszük, hogy a mérési eredmények szórása a mérési eljárástól függ, nem a mért érték várható értékétől.

$X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$  **független normális eloszlású azonos szórású** valószínűségi változók:  $X_i \sim N(m_1, \sigma^2)$ ,  $Y_i \sim N(m_2, \sigma^2)$ , ahol  $m_1, m_2, \sigma$  ismeretlen paraméterek.

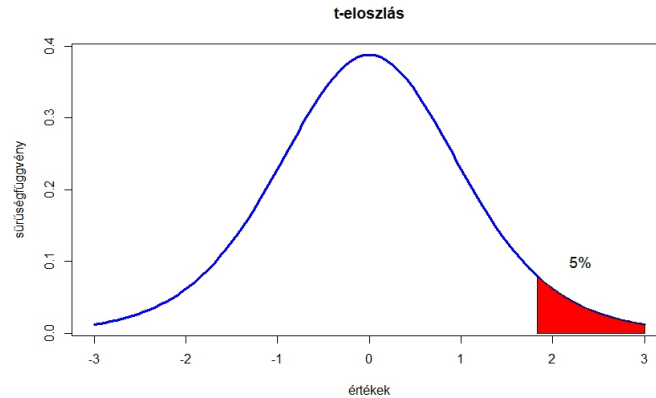
**Egyoldali ellenhipotézis:**  $H_0 : m_1 \leq m_2$ ;  $H_1 : m_1 > m_2$ .

Próbastatisztika (eloszlása  $t$ -eloszlás  $m_1 = m_2$  mellett):

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1 - 1)s_{n_1}^{*2}(X) + (n_2 - 1)s_{n_2}^{*2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}.$$

Ha  $t > \bar{t}_{n_1+n_2-2, \alpha}$ , akkor elvetjük a nullhipotézist, különben elfogadjuk. A  $\bar{t}_{n_1+n_2-2, \alpha}$  kritikus érték az  $f = n_1 + n_2 - 2$  szabadsági fokú **egyoldali  $t$ -próba** kritikus értéke  $\alpha$  szignifikanciaszint mellett (a megfelelő eloszlás  $1 - \alpha$ -kvantilise, 1. ábra)

Ha  $p < \alpha$ : elutasítjuk  $H_0$ -t, az első várható érték szignifikánsan nagyobb a másodiknál.



1. ábra. Az  $f = 9$  szabadsági fokú  $\alpha = 0,05$  szignifikanciaszintű egyoldali  $t$ -próba kritikus értéke:  $\bar{t}_{9,0,05} = 1,83$ .

**$t$ -próba: példa** Egy napon tíz budapesti helyszínen megmérték a  $\text{NO}_2$ -koncentrációt. Az átlag  $352 \mu\text{g}/\text{m}^3$ , a korrigált tapasztalati szórás 8 lett.

- (a) 5%-os szignifikanciaszint mellett elfogadható-e az a hipotézis, hogy a koncentráció a  $350 \mu\text{g}/\text{m}^3$  tájékoztatási küszöbérték alatt van? egymintás egyoldali  $t$ -próba,  $n = 10$ ,  $f = n - 1 = 9$ ,  $\alpha = 0,05$

$$t = \frac{\bar{X} - m_0}{s_n^*} \sqrt{n} = \frac{352 - 350}{8} \sqrt{10} = 0,79 < t_{\text{krit}} = 1,83.$$

elfogadható a nullhipotézis

- (b) Elfogadható-e ugyanez a hipotézis 1%-os szignifikanciaszint mellett?

$t_{\text{krit}} = 2,81 > 0,79$ , elfogadható a hipotézis.

- (c) Londonban 20 mérésből az átlagos koncentráció 376, a korrigált tapasztalati szórás 16 lett.  $\alpha = 0,05$  szignifikanciaszint (terjedelem) mellett állíthatjuk-e, hogy Londonban szignifikánsan nagyobb a  $\text{NO}_2$  koncentrációja?

kétmintás egyoldali  $t$ -próba,  $f = n_1 + n_2 - 2 = 10 + 20 - 2 = 28$ ,  $\alpha = 0,05$ ,  $t_{\text{krit}} = 1,701$ .

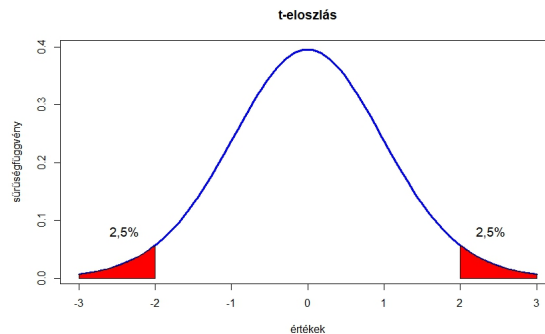
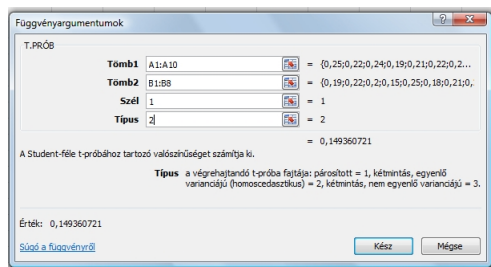
$$t = \frac{\bar{Y} - \bar{X}}{\sqrt{(n_1 - 1)s_{n_1}^{*2}(X) + (n_2 - 2)s_{n_2}^{*2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} = \frac{376 - 352}{\sqrt{9 \cdot 8^2 + 19 \cdot 16^2}} \cdot \sqrt{\frac{10 \cdot 20 \cdot 28}{30}} = 4,44.$$

Londonban szignifikánsan nagyobb a  $\text{NO}_2$  koncentrációja. A  $p$ -érték:  $p = 6,4 \cdot 10^{-5}$ .

Ez a kétmintás egyoldali  $t$ -próba R-ben: `t.test(budapest, london, alternative = "less"), paired = FALSE, var.equal = TRUE)`

## 2. Kétmintás, kétoldali, párosítatlan Student-féle $t$ -próba

A **várható érték összehasonlítására** azonos szórás esetén (two-sample two-sided unpaired Student  $t$ -test).



2. ábra. Az  $f = 29$  szabadsági fokú  $\alpha = 0,05$  szignifikanciaszintű kétoldali  $t$ -próba kritikus értéke:  $t_{29;0,05} = 2,04$ , illetve kétmintás egyoldali Student-féle  $t$ -próba az excelben; a válasz a  $p$ -érték

$X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$  **független normális eloszlású azonos szórású** valószínűségi változók:  $X_i \sim N(m_1, \sigma^2), Y_i \sim N(m_2, \sigma^2)$ , ahol  $m_1, m_2, \sigma$  ismeretlen paraméterek.

**Kétoldali ellenhipotézis:**  $H_0 : m_1 = m_2; H_1 : m_1 \neq m_2$ .

Próbastatisztika (eloszlása  $t$ -eloszlás  $H_0$  mellett):

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1 - 1)s_{n_1}^{*2}(X) + (n_2 - 1)s_{n_2}^{*2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

Ha  $|t| > t_{n_1+n_2-2, 1-\alpha/2}$ , akkor elvetjük a nullhipotézist, különben elfogadjuk. A  $t_{n_1+n_2-2, 1-\alpha/2}$  kritikus érték az  $f = n_1 + n_2 - 2$  szabadsági fokú **kétoldali**  $t$ -próba kritikus értéke  $\alpha$  szignifikanciaszint mellett (a megfelelő eloszlás  $1 - \alpha/2$ -kvantilise: 2. ábra).

Ha  $p < \alpha$ : elutasítjuk  $H_0$ -t, az várható értékek szignifikánsan eltérnek egymástól.

Ez a kétmintás kétoldali  $t$ -próba R-ben: `t.test(x, y, alternative = "equal"), paired = FALSE, var.equal = TRUE)`

### Kétmintás $t$ -próba: példa

Tegyük fel, hogy egy biztosító kétféle termékén az összkár lognormális eloszlású, azaz a logaritmus normális eloszlású (ezután csak a logaritmust tekintjük). Az első termékénél  $n_1 = 20$ , a másodikonál  $n_2 = 12$  éven át figyelték meg az összkár logaritmusát.

Az átlagok és korrigált tapasztalati szórások: ( $X_1, \dots, X_{20}$  az első minta,  $Y_1, \dots, Y_{12}$  a második):

$$\bar{X} = 18,4, \quad s_n^*(X) = 1,2, \quad \bar{Y} = 19,9, \quad s_n^*(Y) = 1,3.$$

Állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy a kétféle termék esetén szignifikánsan eltérő az összkár? Feltételezzük, hogy a minták **függetlenek, normális eloszlásúak, azonos szórásúak**.

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1 - 1)s_{n_1}^{*2}(X) + (n_2 - 1)s_{n_2}^{*2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

Behelyettesítve:

$$t = \frac{18,4 - 19,9}{\sqrt{19 \cdot 1,2^2 + 11 \cdot 1,3^2}} \cdot \sqrt{\frac{20 \cdot 12 \cdot 30}{32}} = -3,3.$$

Az  $f = n_1 + n_2 - 2 = 20 + 12 - 2 = 30$  szabadsági fokú **kétoldali**  $t$ -próba kritikus értéke  $\alpha = 0,05$  szignifikanciaszint mellett:  $t_{30,0,05} = 2,042$ .

Itt  $|t| = 3,3 > 2,042 = t_{30,0,05}$ , ezért **elutasítjuk  $H_0$ -t**. A kétféle összkár logaritmusának várható értéke **szignifikánsan különböző** – ha a szórások azonosak, és a próba alkalmazható (ezt eddig feltettük).

A  $p$ -érték:  $p = 0,0025 < 0,05$ , az eltérés szignifikáns.

### 3. Normális eloszlásra vonatkozó további kétmintás próbák

Az alábbiakat kell ellenőrizni kétmintás próbáknál:

- A minta **normális eloszlású**, vagy a mintaelemszám elég nagy és a szórás feltehetően véges (a centrális határeloszlástétel alapján az átlag közel normális eloszlású).
- Kétmintás esetben: a **két minta egymástól független** ("unpaired" eset). Ha a két minta természetes módon párosítható, **párosított** ("paired") próba alkalmazható. Példa: megfigyeljük húsz ember egyhavi kiadását januárban és áprilisban. Igaz-e, hogy a januári szignifikánsan eltér az áprilistól?
- Ha a **szórásokról feltételezhetjük, hogy megegyeznek**: a Student-féle  $t$ -próba alkalmazható.
- Ha a **szórásokról nem tételezhetjük fel, hogy megegyeznek**: a Welch-féle  $t$ -próba alkalmazható.

#### 3.1. Példa: párosított $t$ -próba

1991 és 2010 között feljegyezték az éves csapadékösszeget Budapesten, illetve Szegeden. Az átlag Budapesten 533 mm, a korrigált tapasztalati szórás 139, Szegeden az átlag 540 mm, a korrigált tapasztalati szórás 143 lett (forrás: Országos Meteorológiai Szolgálat). Állíthatjuk-e, hogy Szegeden szignifikánsan nagyobb a csapadékmennyiség várható értéke?

év	1991	1992	1993	1994	1995	...	átlag	$s_n^*$
Budapest	594	364	505	481	575	...	<b>533</b>	139
Szeged	617	457	408	399	562	...	<b>540</b>	143

A két adatsor **nem független**, mert egy éven belül a két város időjárása nem független (az egyes minták sem teljesen függetlenek, és nem biztos, hogy normális eloszlásúak). Ezért **párosított** (paired)  $t$ -próba alkalmazható, egyoldali nullhipotézissel.

$H_0 : m_1 \geq m_2$ ,  $H_1 : m_1 < m_2$ , ahol  $m_1$  a budapesti,  $m_2$  a szegedi csapadékmennyiség várható értéke.

A próbát elvégezve a  $p$ -értékre 0,366 adódott. Ez több, mint  $\alpha = 0,05$ .

**Elfogadjuk** a nullhipotézist, az adatok alapján Szegeden nem több szignifikánsan a csapadékmennyiség várható értéke, mint Budapesten.

Ez a kétmintás kétoldali  $t$ -próba R-ben: `t.test(x, y, alternative = "equal"), paired = TRUE, var.equal = TRUE)`

### 3.2. Welch-féle $t$ -próba

A **várható érték összehasonlítására** párosítatlan esetben (two-sample two-sided unpaired Welch  $t$ -test). Legyenek  $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$  **független normális eloszlású** valószínűségi változók:  $X_i \sim N(m_1, \sigma_1^2)$ ,  $Y_i \sim N(m_2, \sigma_2^2)$ , ahol  $m_1, m_2, \sigma_1, \sigma_2$  ismeretlen paraméterek.

**Kétoldali ellenhipotézis:**  $H_0 : m_1 = m_2$ ;  $H_1 : m_1 \neq m_2$ .

Próbastatisztika:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_{n_1}^{*2}(X)}{n_1} + \frac{s_{n_2}^{*2}(Y)}{n_2}}}.$$

Ha  $|t| > t_{f, 1-\alpha}$ , akkor elvetjük a nullhipotézist, különben elfogadjuk. A  $t_{f, 1-\alpha}$  kritikus érték az  $f$  szabadsági fokú **kétoldali**  $t$ -próba kritikus értéke  $\alpha$  szignifikanciaszint mellett (a megfelelő eloszlás  $1 - \alpha/2$ -kvantilise).

Szabadsági fok:

$$f \approx \frac{\left(\frac{s_{n_1}^{*2}(X)}{n_1} + \frac{s_{n_2}^{*2}(Y)}{n_2}\right)^2}{\frac{s_{n_1}^{*4}(X)}{n_1^2(n_1-1)} + \frac{s_{n_2}^{*4}(Y)}{n_2^2(n_2-1)}}.$$

Ha  $p < \alpha$ : elutasítjuk  $H_0$ -t, az várható értékek szignifikánsan eltérnek egymástól.

Ez a kétmintás kétoldali  $t$ -próba R-ben: `t.test(x, y, alternative = "equal"), paired = FALSE, var.equal = FALSE)`

**Egyoldali ellenhipotézis:**  $H_0 : m_1 \leq m_2$ ;  $H_1 : m_1 > m_2$ .

Próbastatisztika: ugyanaz, mint az előbb

Ha  $t > \bar{t}_{f, 1-\alpha}$ , akkor elvetjük a nullhipotézist, különben elfogadjuk. A  $\bar{t}_{f, 1-\alpha}$  kritikus érték az  $f$  szabadsági fokú **egyoldali**  $t$ -próba kritikus értéke  $\alpha$  szignifikanciaszint mellett (a megfelelő eloszlás  $1 - \alpha$ -kvantilise).

Szabadsági fok: ugyanúgy, mint az előbb.

Ha  $p < \alpha$ : elutasítjuk  $H_0$ -t, az várható értékek szignifikánsan eltérnek egymástól.

Ez a kétmintás egyoldali  $t$ -próba R-ben: `t.test(x, y, alternative = "greater"), paired = FALSE, var.equal = FALSE)`

### 4. $F$ -próba

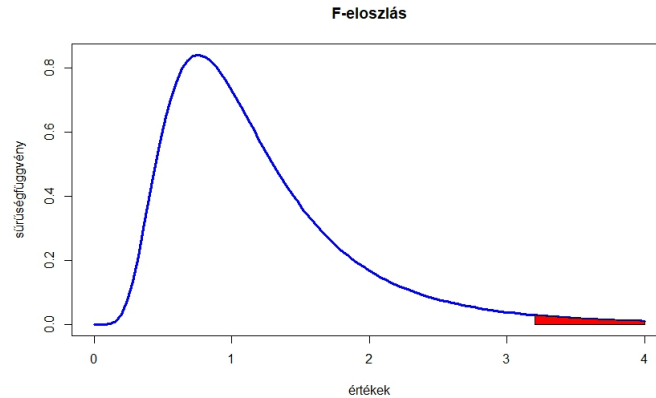
**Független** normális eloszlású minták **szórásának** összehasonlítására. Ugyanis az is előfordulhat, hogy nem (csak) a várható értéket, hanem a szórást is össze akarjuk hasonlítani. Például: igaz-e, hogy a havi kiadások szórása nagyobb Budapesten, mint más településeken, vagy hogy két mérési eljárás közül az egyiknek szignifikánsan nagyobb a szórása.

Emlékeztetőül: Legyenek  $m, n$  pozitív egészek,  $X_1, \dots, X_m, Y_1, Y_2, \dots, Y_n$  pedig független standard normális eloszlású valószínűségi változók. Ekkor az

$$F = \frac{n(X_1^2 + X_2^2 + \dots + X_m^2)}{m(Y_1^2 + Y_2^2 + \dots + Y_n^2)}$$

valószínűségi változó eloszlását  $m, n$  szabadsági fokú  **$F$ -eloszlásnak** nevezzük.

- Legyenek most  $X_1, X_2, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$  független normális eloszlású valószínűségi változók, ahol  $X_i \sim N(m_1, \sigma_1^2)$ ,  $Y_i \sim N(m_2, \sigma_2^2)$ . Itt  $m_1, m_2, \sigma_1, \sigma_2$  ismeretlen paraméterek.



3. ábra. Az  $F$ -próba kritikus értéke:  $F_{19,11} = 3,24$ , ez az eloszlás  $1 - \alpha/2 = 0,975$ -kvantilise

- Kétoldali ellenhipotézis:  $H_0 : \sigma_1 = \sigma_2$ ;  $H_1 : \sigma_1 \neq \sigma_2$ .
- Próbastatisztika (eloszlása  $F$ -eloszlás  $H_0$  mellett):

$$F = \frac{s_{n_1}^{*2}}{s_{n_2}^{*2}}.$$

- Ha  $F > F_{n_1-1, n_2-1}$  vagy  $1/F > F_{n_2-1, n_1-1}$ , akkor elvetjük a nullhipotézist, különben elfogadjuk, ahol  $F_{f_1, f_2}$  az  $f_1, f_2$  szabadsági fokú az  $F$ -eloszlás  $1 - \alpha/2$ -kvantilise.

$p < 0,05$ : a szórások szignifikánsan eltérnek.

### Kétmintás $F$ -próba: példa

A korábbi példában az összkár logaritmusának szórását szeretnénk összehasonlítani. Az első termék esetében  $n_1 = 20$ , a második esetében  $n_2 = 12$  éven át figyelték meg az összkárt. Az átlagok és korrigált tapasztalati szórások ( $X_1, \dots, X_{20}$  az első minta,  $Y_1, \dots, Y_{12}$  a második):

$$\bar{X} = 18,4, \quad s_n^*(X) = 1,2, \quad \bar{Y} = 19,9, \quad s_n^*(Y) = 1,3.$$

Állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy a kétféle termék esetében szignifikánsan eltérő az összkár szórása? Feltételezzük, hogy a minták **függetlenek, normális eloszlásúak**.

$$H_0 : \sigma_1 = \sigma_2, \quad H_1 : \sigma_1 \neq \sigma_2$$

A próbastatisztika értéke:  $F = \frac{s_{n_1}^{*2}}{s_{n_2}^{*2}} = \frac{1,2^2}{1,3^2} = 0,85$ , és  $\frac{1}{F} = \frac{1,3^2}{1,2^2} = 1,17$ .

Az  $(f_1, f_2) = (n_1 - 1, n_2 - 1) = (19, 11)$  szabadsági fokú  $F$ -próba kritikus értéke  $\alpha = 0,05$  esetén:  $3,24$ , míg az  $(f_2, f_1) = (n_2 - 1, n_1 - 1) = (11, 19)$  szabadsági fok esetén  $2,76$ .

Mivel  $F < 3,24$  és  $1/F < 2,76$ , **elfogadjuk a nullhipotézist**, a szórások nem térnek el szignifikánsan. (Vagyis, a fent alkalmazott kétmintás  $t$ -próbánál nem volt teljesen helytelen a feltételezés, hogy a szórások megegyeznek, de általában, ha nem biztos, hogy a szórások lényegében azonosak, akkor inkább a nem feltétlenül egyenlő szórásokra vonatkozó eljárást érdemes használni).

## 5. Illeszkedésvizsgálat

Az alábbi eljárással azt tudjuk ellenőrizni, hogy bizonyos események valószínűsége, vagy egy diszkrét valószínűségi változó eloszlása közelítőleg megegyezik-e az általunk alkotott elképzeléssel,

vagy az adatok alapján azt állíthatjuk, hogy a valódi valószínűségek szignifikánsan eltérnek az előzetesen megadottól.

Például: egy politikai elemző azt állítja, hogy a pártot választók között az  $A$  párt támogatottsága 40%, a  $B$  párté 20%, a  $C$  párté 15%, a többiek pedig kisebb pártok valamelyikére szavaznak. Megkérdezzük 200, függetlenül választott szavazót (aki részt venne a választáson és érvényesen szavazna). Közülük 92-en az  $A$  pártot, 38-an a  $B$ -t, 31-en a  $C$ -t támogatnák szavazatukkal, a többiek a kisebb pártok valamelyikét. Ez alapján  $\alpha = 0,05$  szignifikanciaszint (elsőfajú hiba- valószínűség, terjedelem) mellett elfogadható-e az elemző állítása?

Tekintsük az alábbi eseményeket:

$A$ : egy véletlenszerűen választott szavazó az  $A$  pártot támogatja  
 $B$ : egy véletlenszerűen választott szavazó a  $B$  pártot támogatja  
 $C$ : egy véletlenszerűen választott szavazó a  $C$  pártot támogatja  
 $D$ : egy véletlenszerűen választott szavazó a kisebb pártok valamelyikét támogatja

A feltételezésünk szerint ezek közül az események közül, egy szavazót kiválasztva, pontosan az egyik következik be, vagyis  $A, B, C, D$  teljes eseményrendszert alkotnak (uniójuk az összes lehetőség halmaza,  $\Omega$ , páronkénti metszeteik üresek).

A nullhipotézisben minden eseményhez egy valószínűség tartozott, úgy, hogy a valószínűségek összege 1 (annak megfelelően, hogy pontosan az egyik esemény következik be).

$H_0 : \mathbb{P}(A) = 40\%, \mathbb{P}(B) = 20\%, \mathbb{P}(C) = 15\%, \mathbb{P}(D) = 25\%$ .

$H_1$ : a nullhipotézisben megadott feltételek közül legalább az egyik nem teljesül.

Általánosabban: legyen  $A_1, A_2, \dots, A_r$  teljes eseményrendszer (olyan események, amik közül pontosan az egyik következik be, azaz uniójuk a teljes eseménytér, páronkénti metszetük üres),  $p_1, p_2, \dots, p_r$  pedig olyan nemnegatív számok, melyek összege 1.

$H_0 : \mathbb{P}(A_k) = p_k$  minden  $k = 1, 2, \dots, r$ -re.

$H_1 : \mathbb{P}(A_k) \neq p_k$  valamelyik  $k = 1, 2, \dots, r$ -re.

Ezekben a feladatokban  $\chi^2$ -próbát, azon belül illeszkedésvizsgálatot végezhetünk.

- $n$  független megfigyelést végzünk.
- $N_k$ : hányszor következett be  $A_k$ , vagyis az  $A_k$  gyakorisága.
- Ha van  $k$ , hogy  $N_k < 5$ : néhány osztályt össze kell vonnunk, hogy a próbát alkalmazhassuk (vagyis  $A_j$  és  $A_k$  helyett  $A_j \cup A_k$ -t és  $p_j + p_k$ -t tekintjük, és ezután végezzük el a tesztet).
- Próbastatisztika:

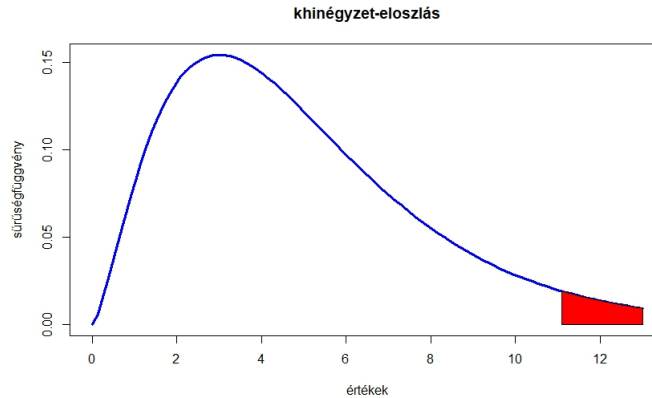
$$\chi^2 = \sum_{k=1}^r \frac{(N_k - n \cdot p_k)^2}{n \cdot p_k}.$$

Ez minél nagyobb, annál nagyobb az eltérés a nullhipotézistől. Hiszen egyrészt  $H_0$  esetén  $N_k$  várható értéke  $np_k$ . Másrészt a „várt” és a „megfigyelt” gyakoriság közötti eltérés annál jobban számít, minél kisebb a várt érték, arányaiban annál nagyobb a különbség.

**5.1. Állítás.**  $A H_0$  nullhipotézis teljesülése esetén

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \sum_{k=1}^r \frac{(N_k - n \cdot p_k)^2}{n \cdot p_k} \leq t \right) = \mathbb{P}(U_{r-1} \leq t)$$

teljesül minden  $t$ -re, ahol  $U_{r-1}$  eloszlása  $r - 1$  szabadsági fokú  $\chi^2$ -eloszlás, azaz eloszlása megegyezik  $Z_1^2 + Z_2^2 + \dots + Z_{r-1}^2$  eloszlásával, ahol  $Z_1, Z_2, \dots, Z_{r-1}$  független standard normális eloszlású valószínűségi változók.



4. ábra. Az  $f = 5$  szabadsági fokú  $\chi^2$ -eloszlás sűrűségfüggvénye. Az  $\alpha = 0,05$  szignifikanciaszintű próba kritikus értéke:  $c_{\text{krit}} = 11,1$ .

A célunk egy olyan eljárás, amivel a nullhipotézis téves elutasításának valószínűsége legfeljebb  $\alpha$ . A  $\chi^2$  a nullhipotézistől való eltérést mutatja, vagyis akkor utasítjuk el  $H_0$ -t, ha  $\chi^2$  értéke nagyobb egy kritikus értéknél. Ezt pedig úgy választjuk, hogy annak valószínűsége, hogy  $H_0$  mellett  $\chi^2 > c_{\text{krit}}$  legyen, legyen éppen  $\alpha$ .

Ezért legyen  $c_{\text{krit}}$  az  $f = r - 1$  szabadsági fokú  $\chi^2$ -próba kritikus értéke  $\alpha$  szignifikanciaszint mellett, vagyis az  $f = r - 1$  szabadsági fokú  $\chi^2$ -eloszlás  $1 - \alpha$  kvantilise (4. ábra).

$\chi^2 > c_{\text{krit}}$  vagy  $p < \alpha$ : elutasítjuk  $H_0$ -t, az eloszlás **szignifikánsan eltér** ( $p_k$ )-től.

$\chi^2 \leq c_{\text{krit}}$  vagy  $p \geq \alpha$ : elfogadjuk  $H_0$ -t, az eloszlás **nem tér el szignifikánsan** ( $p_k$ )-től.

Az a feltétel, hogy minden  $N_k$  legyen legalább 5, abból adódik, hogy csak a valószínűség li-meszeről tudjuk, hogy megegyezik a  $\chi^2$ -eloszlásból számolt valószínűséggel, tehát véges mintaelemszám esetén legfeljebb csak közelítésről van szó, és ezért **túl kicsi mintaelemszám esetén a próba nem alkalmazható**. Ugyanakkor **túl nagy mintaelemszám esetén a próba túl érzékennyé válik**, például egy 20000 méretű mintából be lehet látni, hogy vasárnap szignifikánsan gyakoribbak a nagyobb földrengések, mint más napokon, ami nem ennek az állításnak az igazságát, hanem a próba túlzott érzékenységét mutatja.

A  $\chi^2$ -próbában a  **$p$ -érték**

- ahogy általában is, a legnagyobb olyan szignifikanciaszint, ami mellett a nullhipotézist elfogadjuk;
- azaz  $p < \alpha$  esetén elutasítjuk a nullhipotézist, különben elfogadjuk;
- tehát az a kérdés, hogy milyen  $\alpha$ -ra lenne igaz, hogy  $\chi^2$  éppen megegyezik a kritikus értékkel;
- ez tehát annak valószínűsége, hogy az  $r - 1$  szabadsági fokú  $\chi^2$ -eloszlás legalább  $\chi^2$ ;
- másképpen:  $p = \mathbb{P}(U_{r-1} \geq \chi^2)$ , ahol  $U_{r-1}$  eloszlása  $r - 1$  szabadsági fokú  $\chi^2$ -eloszlás;
- a 4. ábrához hasonlóan, ez a  $\chi^2$  értékétől jobbra eső terület lenne;
- kiszámítás az R-ben, ha  $\chi^2 = s$ : `pchisq(s, df=r-1, lower.tail=FALSE)` (valószínűséget számolunk, df a szabadsági fok, és annak valószínűsége kell, hogy  $s$ -nél nagyobb az érték, ezt állítja az utolsó paraméter, enélkül a balra lévő területet kapnánk)
- ahogy általában is, minél kisebb a  $p$ -érték, annál szignifikánsabb az eltérés.

**Illeszkedésvizsgálat: példa.** Tekintsük a fent megfogalmazott példát a pártok támogatottságáról.

$H_0 : \mathbb{P}(A) = 40\%, \mathbb{P}(B) = 20\%, \mathbb{P}(C) = 15\%, \mathbb{P}(D) = 25\%$ .

$H_1$ : a nullhipotézisben megadott feltételek közül legalább az egyik nem teljesül.

Itt  $N_1 = 92, N_2 = 38, N_3 = 31, N_4 = (200 - 92 - 38 - 31) = 39$ .

Minden osztályba esik legalább 5 megfigyelés, nem túl kicsi a mintaelemszám, nem kell osztályokat összevonni, és a 200 még talán nem is számít túl soknak.

$$\chi^2 = \sum_{k=1}^r \frac{(N_k - n \cdot p_k)^2}{n \cdot p_k} = \frac{(92 - 200 \cdot 0,4)^2}{200 \cdot 0,4} + \frac{(38 - 200 \cdot 0,2)^2}{200 \cdot 0,2} + \frac{(31 - 200 \cdot 0,15)^2}{200 \cdot 0,15} + \frac{(39 - 200 \cdot 0,25)^2}{200 \cdot 0,25} = 4,35.$$

A próba szabadsági foka  $f = r - 1 = 3$ . A kritikus érték (táblázatból, vagy `qchisq(0.95, df=3)` az R-ben):  $c_{\text{krit}} = 7,81$ , ha  $\alpha = 0,05$ . Mivel  $\chi^2 < c_{\text{krit}}$ , **elfogadjuk a nullhipotézist**, az adatok nem mutatnak szignifikáns eltérést az elemző állításától.

Megvalósítás az R-ben:

```
> adat=c(92, 38, 31, 39)
> val=c(0.4, 0.2, 0.15, 0.25)
> chisq.test(adat, p=val)
```

Chi-squared test for given probabilities

data: adat

X-squared = 4.3533, df = 3, p-value = **0.2258**

A próba tulajdonságai:

- a  $\chi^2$ -próba **aszimptotikus próba**, vagyis a próbastatisztika eloszlása nem pontosan  $\chi^2$ -eloszlás, csak ahhoz tart, ha a mintaelemszámmal végtelenhez tartunk
- emiatt: minden különálló csoportba kell esnie **legalább négy** (vagy inkább hat) megfigyelésnek, ez biztosítja az elég nagy mintaelemszámot
- ugyanakkor: túl **nagy mintaelemszámmal** a  $\chi^2$ -próba **túlságosan érzékeny**, túl gyakran mutat ki szignifikáns eltérést (például egy 20000 elemű mintából be lehet látni, hogy vasárnap szignifikánsan gyakrabban vannak földrengések, mint a többi napon – ez túl nagy mintaelemszám, érdemes egy kisebb mintát venni a nagy adathalmazból)

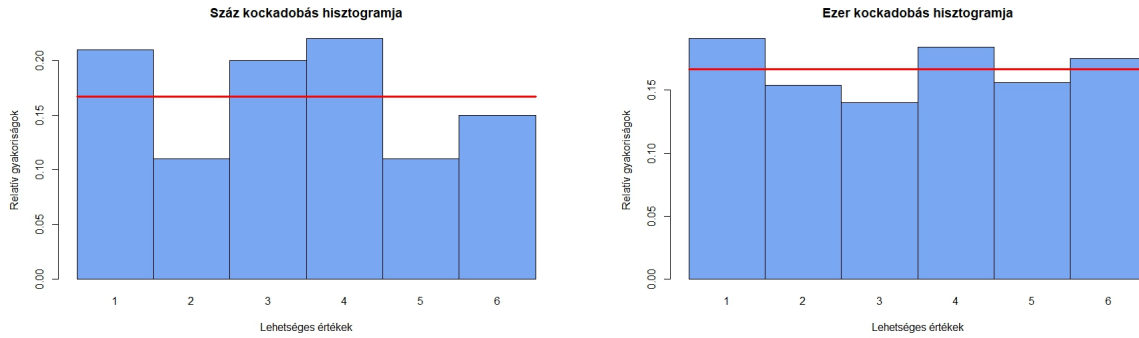
**Illeszkedésvizsgálat: példa**

Dobókockával dobunk százszor (5. ábra). A terjedelmet  $\alpha = 0,05$ -nek választva elfogadható-e, hogy szabályos a dobókocka?

érték	1	2	3	4	5	6
gyakoriság	21	11	20	22	11	15

Minden szám legalább négyszer előfordult, alkalmazhatjuk a  $\chi^2$ -próbát.  $A_i$ :  $i$ -t dobunk,  $r = 6$ ,  $p_k = 1/6, k = 1, 2, \dots, 6$ .

$H_0 : \mathbb{P}(A_k) = 1/6$  minden  $k$ -ra;  $H_1 : \mathbb{P}(A_k) \neq 1/6$  valamelyik  $k$ -ra



5. ábra. Száz, illetve ezer kockadobás histogramja. Elfogadható-e, hogy szabályos a dobókocka?

$$\chi^2 = \sum_{k=1}^r \frac{(N_k - n \cdot p_k)^2}{n \cdot p_k} = \frac{(21 - 100 \cdot 1/6)^2}{100 \cdot 1/6} + \frac{(11 - 100 \cdot 1/6)^2}{100 \cdot 1/6} + \dots + \frac{(15 - 100 \cdot 1/6)^2}{100 \cdot 1/6} = 7,52.$$

$\chi^2 = 7,52 < c_{\text{krit}} = 11,1$ , illetve a  $p$ -értékre  $0,1847 > 0,05$ .

Elfogadjuk  $H_0$ -t, elfogadható, hogy a dobókocka szabályos, **nincs szignifikáns eltérés** az egyenletes eloszlástól.

Egy másik adatsor: dobókockával dobunk ezerszer (5. ábra). A terjedelmet  $\alpha = 0,05$ -nek választva elfogadható-e, hogy szabályos a dobókocka?

érték	1	2	3	4	5	6
gyakoriság	191	154	140	184	156	175

$H_0 : \mathbb{P}(A_k) = 1/6$  minden  $k$ -ra;  $H_1 : \mathbb{P}(A_k) \neq 1/6$  valamelyik  $k$ -ra

$$\chi^2 = 11,68; \quad f = r - 1 = 5; \quad \alpha = 0,05; \quad c_{\text{krit}} = 11,1$$

$\chi^2 = 11,68 > c_{\text{krit}} = 11,1$ , illetve a  $p$ -értékre  $0,039 < 0,05$ .

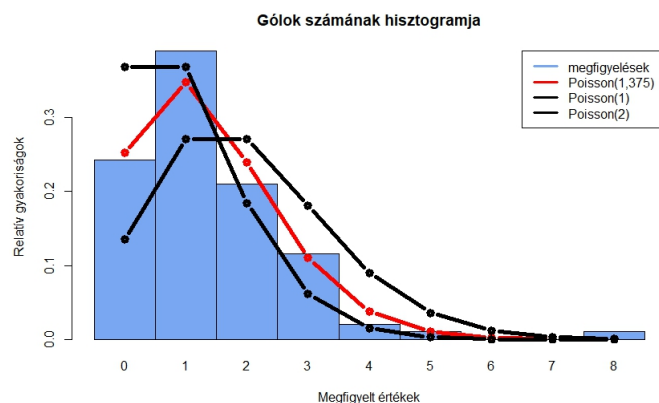
**Elutasítjuk  $H_0$ -t**, nem fogadható el, hogy a dobókocka szabályos, a minta alapján az eloszlás **szignifikánsan eltér** az egyenletes eloszlástól.

## 6. Becsléses illeszkedésvizsgálat

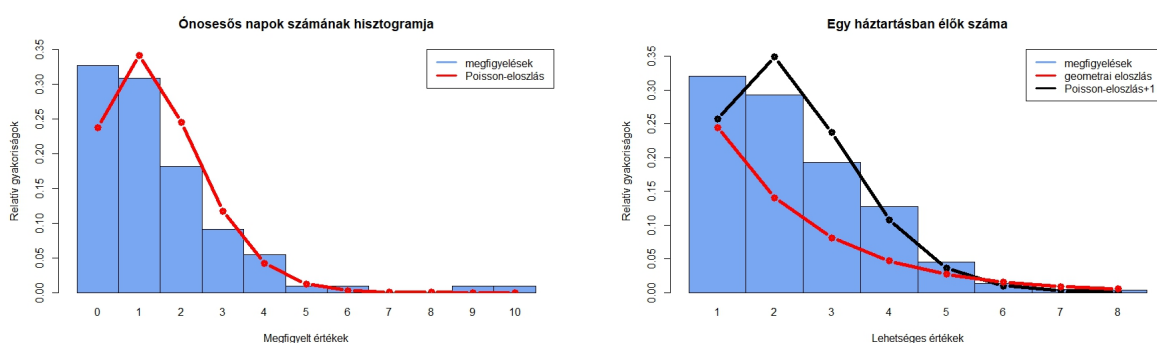
Gyakran előfordul, hogy az eloszlásról nem egy pontos valószínűségekkel leírható hipotézisünk van, hanem csak az, hogy valamilyen eloszláscsaládból származik, például Poisson-eloszlású (ennek a folytonos változata, amikor például az a kérdés, hogy egy eloszlás normális eloszlású-e, erről később lesz szó). A fenti  $\chi^2$ -próban alapuló illeszkedésvizsgálat egy módosított változata a diszkrét eloszlások esetén alkalmazható.

Elfogadható-e 0,05 terjedelem (szignifikanciaszint) mellett, hogy az egy futballmérkőzésen lőtt gólok száma Poisson-eloszlású?

A 6. ábrán láthatók megfigyelt adatok  $n = 95$  elemű mintából, melyek átlaga  $\bar{X} = 1,379$ , és a  $\hat{\lambda} = 1,379$  paraméterű Poisson-eloszlás:  $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ .



6. ábra. A gólok számának hisztogramja és néhány különböző paraméterű Poisson-eloszlás



7. ábra. Az ónosesős napok számának hisztogramja és a  $\hat{\lambda} = 1,44$  paraméterű Poisson-eloszlás; gy háztartásban élők számának hisztogramja (KSH, 2011), Poisson-eloszlás és geometriai eloszlás

Elfogadható-e 0,05 szignifikanciaszint mellett, hogy Budapesten az ónosesős napok száma egy év alatt Poisson-eloszlású?

A 7. ábrán láthatók megfigyelt adatok  $n = 110$  elemű mintából (1901–2010, Országos Meteorológiai Szolgálat), melyek átlaga  $\bar{X} = 1,44$ , és a  $\hat{\lambda} = 1,44$  paraméterű Poisson-eloszlás:  $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ .

Egy másik példa a 7. ábrán látható: ez az egy háztartásban élők számának hisztogramja (forrás: KSH, 2011), és a geometriai eloszlás ( $p = 1/\bar{X}$ ), illetve a  $\text{Poisson}(\bar{X})$ -eloszlás egyvel eltolva. Itt  $\bar{X} = 2,36$  az átlag, és  $n = 4105698$  a háztartások száma, **túl nagy a mintaelemszám**.

### 6.1. A $\chi^2$ -próba alkalmazása

Az illeszkedésvizsgálathoz hasonlóan legyen  $A_1, A_2, \dots, A_r$  teljes eseményrendszer, azaz olyan események, amik közül pontosan az egyik következik be.  $N_k$ : hányszor következik be  $A_k$  egy  $n$  elemű független mintában. Feltesszük, hogy  $N_k \geq 5$  minden  $k$ -ra, ha nem, osztályokat vonunk össze. Adott  $p_k(\lambda)$  minden  $\lambda \in \mathcal{L}$ -re.

$H_0$ : van olyan  $\lambda \in \mathcal{L}$ , melyre  $\mathbb{P}(A_k) = p_k(\lambda)$  minden  $k = 1, 2, \dots, r$ -re.

$H_1$ : nincs ilyen  $\lambda \in \mathcal{L}$ , az eloszlás **szignifikánsan eltér** a  $(p_k(\lambda))$  eloszláscsaládtól.

A  $\lambda$  paramétervektor **maximumlikelihood-becslése** legyen  $\hat{\lambda}$ , és legyen  $\hat{p}_k = p_k(\hat{\lambda})$ . A  $\lambda$

dimenziója, vagyis a becült paraméterek száma  $d$ . Próbastatisztika:

$$\chi^2 = \sum_{k=1}^r \frac{(N_k - n \cdot \hat{p}_k)^2}{n \cdot \hat{p}_k}.$$

Legyen  $f = r - d - 1$ , és  $c_{\text{krit}}$  az  $f$  szabadsági fokú  $\chi^2$ -próba kritikus értéke  $\alpha$  szignifikanciaszint mellett (**a szabadsági fokból levonjuk a becült paraméterek számát**).  $H_0$ -t elutasítjuk, ha  $\chi^2 > c_{\text{krit}}$  (azaz  $p < \alpha$ ), ilyenkor a minta szignifikánsan eltér a nullhipotézisben szereplő eloszláscsaládtól. Ha  $\chi^2 \leq c_{\text{krit}}$ , akkor elfogadjuk a nullhipotézist.

A  $p$ -érték az illeszkedésvizsgálathoz hasonlóan számolható, ez annak valószínűsége, hogy az  $f = r - d - 1$  szabadsági fokú  $\chi^2$ -eloszlás több-e a fent kiszámított  $\chi^2$ -nél. Az  $\alpha$ -nál kisebb  $p$ -érték jelenti a nullhipotézis elutasítását.

## 6.2. Becsléses illeszkedésvizsgálat: példa

Az egy futballmérkőzésen lőtt gólok száma a világbajnokság  $n = 95$  mérkőzésén (6. ábra):

gólok száma	0	1	2	3	4	5	6	7	8
mérkőzések száma	23	37	20	11	2	1	0	0	1

Poisson-esetben a  $\lambda$  paraméter maximumlikelihood-becslése:

$$\hat{\lambda} = \bar{X} = \frac{0 \cdot 23 + 1 \cdot 37 + 2 \cdot 20 + 3 \cdot 11 + 4 \cdot 2 + 5 \cdot 1 + 8 \cdot 1}{95} = 1,379.$$

Mivel vannak olyan osztályok, ahova 5-nél kevesebb megfigyelés esik, a beosztást módosítjuk (viszont most kivételesen megelégszünk a legalább 4 megfigyeléssel osztálynként):

gólok száma	0	1	2	3	$\geq 4$
mérkőzések száma	23	37	20	11	4

$H_0$ : az eloszlás **Poisson-eloszlásból** származik valamely  $\lambda > 0$ -val.

$H_1$ : az eloszlás **eltér a Poisson-eloszlástól**.

$\hat{\lambda} = 1,379$  a paraméter maximumlikelihood-becslése. Ekkor

$$\hat{p}_k = \frac{\hat{\lambda}^k}{k!} e^{-\hat{\lambda}} \quad (k = 0, 1, 2, \dots)$$

a Poisson-eloszlás definíciójába a  $\hat{\lambda}$  becült paramétert helyettesítve.

gólok száma	0	1	2	3	$\geq 4$
mérkőzések száma	23	37	20	11	4
$n\hat{p}_k$ (Poisson( $\hat{\lambda}$ ))	23,92	32,99	22,75	10,46	4,88

$$\chi^2 = \sum_{k=1}^r \frac{(N_k - n \cdot \hat{p}_k)^2}{n \cdot \hat{p}_k} = \frac{(23 - 23,92)^2}{23,92} + \frac{(37 - 32,99)^2}{32,99} + \dots = 1,04.$$

$$\chi^2 = 1,04; \quad \mathbf{f = r - d - 1} = 5 - 1 - 1 = 3; \quad \alpha = 0,05; \quad c_{\text{krit}} = 7,81.$$

$\chi^2 = 1,04 < 7,81 = c_{\text{krit}}$ , ezért elfogadjuk, hogy a minta Poisson-eloszlású, **nincs szignifikáns eltérés** a Poisson-eloszlástól. A  $p$ -érték:  $p = 0,21$ .

### 6.3. Becsléses illeszkedésvizsgálat: második példa

Az ónosesős napok évenkénti száma  $n = 110$  éven keresztül Budapesten:

ónosesős napok száma	0	1	2	3	4	5	6	7	8	9	10
évek száma	36	34	20	10	6	1	1	0	0	1	1

Poisson-esetben a  $\lambda$  paraméter maximumlikelihood-becslése:

$$\hat{\lambda} = \bar{X} = \frac{0 \cdot 36 + 1 \cdot 34 + 2 \cdot 20 + 3 \cdot 10 + \dots + 10 \cdot 1}{110} = 1,436.$$

Mivel vannak olyan osztályok, ahova 5-nél kevesebb megfigyelés esik, a beosztást módosítjuk (de most is öt helyett négygel megelégszünk):

ónosesős napok száma	0	1	2	3	4	$\geq 5$
évek száma	36	34	20	10	6	4

$H_0$ : az eloszlás **Poisson-eloszlásból** származik valamely  $\lambda > 0$ -val.

$H_1$ : az eloszlás **eltér a Poisson-eloszlástól**.

$\hat{\lambda} = 1,436$  a paraméter maximumlikelihood-becslése. Ekkor

$$\hat{p}_k = \frac{\hat{\lambda}^k}{k!} e^{-\hat{\lambda}} \quad (i = 0, 1, 2, \dots)$$

a Poisson-eloszlás definíciójába a  $\hat{\lambda}$  becült paramétert helyettesítve.

ónosesős napok száma	0	1	2	3	4	$\geq 5$
évek száma	36	34	20	10	6	4
$n\hat{p}_k$ (Poisson( $\hat{\lambda}$ ))	26,17	37,58	26,98	12,91	4,64	1,73

$$\chi^2 = \sum_{k=1}^r \frac{(N_k - n \cdot \hat{p}_k)^2}{n \cdot \hat{p}_k} = \frac{(36 - 26,17)^2}{26,17} + \frac{(34 - 37,58)^2}{37,58} + \dots = 9,88.$$

$$\chi^2 = 9,88; \quad \mathbf{f} = \mathbf{r} - \mathbf{d} - \mathbf{1} = 6 - 1 - 1 = 4; \quad \alpha = 0,05; \quad c_{\text{krit}} = 9,49.$$

$\chi^2 = 9,88 > 9,49 = c_{\text{krit}}$ , ezért elutasítjuk, hogy a minta Poisson-eloszlású, az eloszlás **szignifikánsan eltér** a Poisson-eloszlástól. A  $p$ -érték:  $p = 0,04$ .

**Házi feladat április 21., szerda, 9:00-ig** Állíthatjuk-e, hogy az utóbbi egy hónapban nézett sorozatok száma, illetve az olvasott könyvek száma szignifikánsan eltér a Poisson-eloszlástól? A kettő közül melyik mintára illeszkedik „jobban” a Poisson-eloszlás? Készítsünk ábrát is erről, ahol összehasonlítjuk a hisztogramot és a Poisson-eloszlást a megfelelően becült paraméterrel.