

## 1. 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 hibavalószínűség, 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:

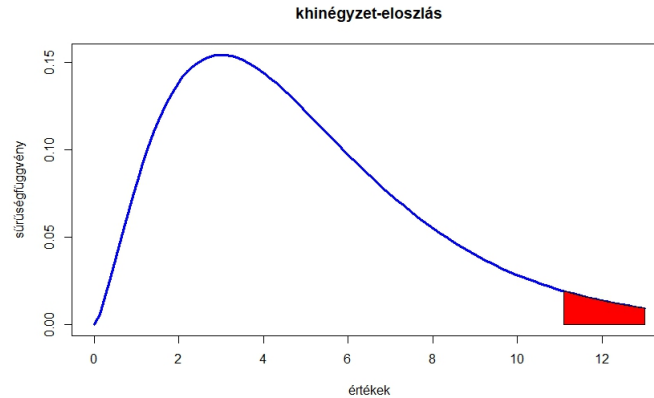
$$\chi^2 = \sum_{k=1}^r \frac{(N_k - n \cdot p_k)^2}{n \cdot p_k}.$$

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.

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

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

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.



1. 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 (1. 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 limeszéről 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 érzékenyvé 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  $\chi^2$ -próbaiban 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;
- az 1. 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.

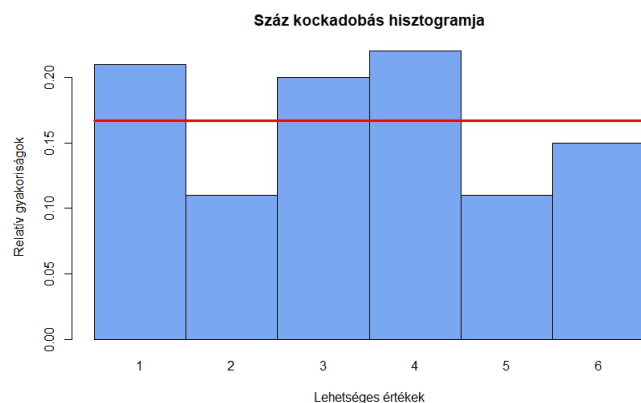
**Illeszkedésvizsgálat: példa.** Tekintsük a fent megfogalmazott példát a pártok támogatottságáról.

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

$H_1$ : 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.



2. ábra. Száz kockadobás hisztogramja. Elfogadható-e, hogy minden szám egyformán valószínű?

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

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**

### Illeszkedésvizsgálat: példa

Dobókockával dobunk százszor (2. 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

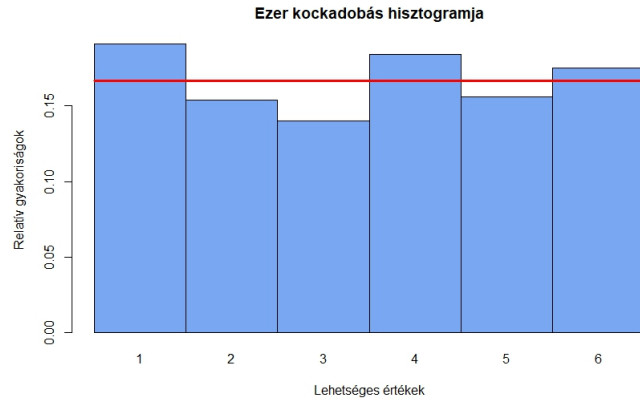
Minden szám legalább négyszer előfordult, alkalmazhatjuk a  $\chi^2$ -próbát.  $A_i$ :  $i$ -t dobunk,  $r = 6$ ,  $p_k = 1/6$ ,  $k = 1, 2, \dots, 6$ .

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

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

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

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



3. ábra. Ezer kockadobás hisztogramja. Elfogadható-e, hogy szabályos a dobókocka?

Egy másik adatsor: dobókockával dobunk ezerszer (3. 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.

## 2. 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  $\chi^2$ -próbán alapuló illeszkedésvizsgálat egy módosított változata a diszkrét eloszlások esetén alkalmazható.

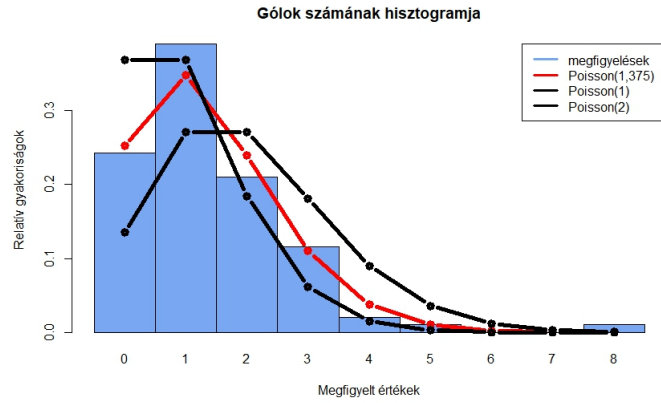
Elfogadható-e 0,05 terjedelem (szignifikanciaszint) mellett, hogy az egy futballmérkőzésen lőtt gólok száma Poisson-eloszlású?

A 4. ábrán láthatók megfigyelt adatok  $n = 95$  elemű mintából, melyek átlaga  $\bar{X} = 1,379$ , és a  $\hat{\lambda} = 1,379$  paraméterű Poisson-eloszlás:  $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ .

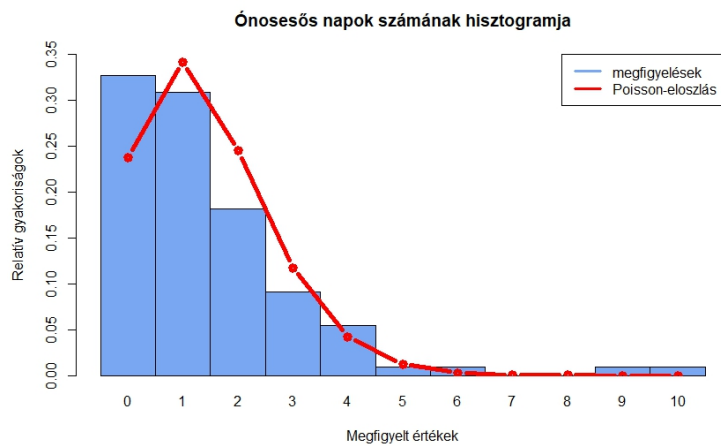
Elfogadható-e 0,05 szignifikanciaszint mellett, hogy Budapesten az ónosos napok száma egy év alatt Poisson-eloszlású?

Az 5. 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:  $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ .

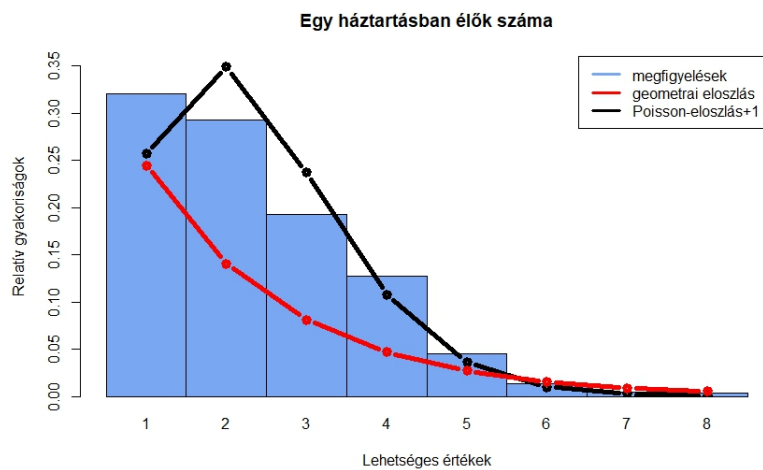
Egy másik példa a 6. ábrán látható: ez az egy háztartásban élők számának hisztogramja (forrás: KSH, 2011), és a geometriai eloszlás ( $p = 1/\bar{X}$ ), illetve a Poisson( $\bar{X}$ )-eloszlás eggyel eltolva. Itt  $\bar{X} = 2,36$  az átlag, és  $n = 4105698$  a háztartások száma, **túl nagy a mintaelemszám**.



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



5. ábra. Az ónosos napok számának hisztogramja és a  $\hat{\lambda} = 1,44$  paraméterű Poisson-eloszlás



6. ábra. Egy háztartásban élők számának hisztogramja (KSH, 2011), Poisson-eloszlás és geometriai eloszlás

## 2.1. 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 becsült 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 becsült 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.

## 2.2. Becsléses illeszkedésvizsgálat: példa

Az egy futballmérkőzésen lőtt gólok száma a világbajnokság  $n = 95$  mérkőzésén (4. 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 osztályonként):

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  $\hat{\lambda}$  becsült 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:  $p = 0,21$ .

### 2.3. Becsléses illeszkedésvizsgálat: második példa

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

ónosos 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):

ónosos 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  $\hat{\lambda}$  becslést paraméter helyettesítve.

ónosos 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

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

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

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

### 2.4. Függetlenségvizsgálat

Ez az eljárás annak eldöntésére szolgál, hogy két szempont szerinti osztályba sorolás független-e egymástól. Például: egy véletlenszerűen választott embert iskolai végzettség és jövedelmi kategória szerint osztályokba sorolva független-e a két szempont.

Két szempont szerint soroljuk osztályokba a megfigyeléseket.

Első szempont:  $A_1, \dots, A_r$  (teljes eseményrendszer, pontosan az egyik következik be, például: iskolai végzettség szerinti kategóriák).

Második szempont:  $B_1, \dots, B_s$  (ez egy másik teljes eseményrendszer, például: jövedelmi kategóriák).

$H_0$ : **a két szempont független** egymástól, azaz  $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(B_j)$  minden  $i, j$ -re.

$H_1$ : a nullhipotézis nem igaz, a két szempont között **összefüggés** van.

$N_{ij}$ : hány olyan megfigyelés van, melyre  $A_i$  és  $B_j$  teljesül.

$N_{i\cdot} = \sum_{j=1}^s N_{ij}$  (azaz az  $A_i$  gyakorisága);  $N_{\cdot j} = \sum_{i=1}^r N_{ij}$  (azaz  $B_j$  gyakorisága);  $n$  pedig az összes megfigyelés száma. Ekkor a próbatisztika, mely  $H_0$  mellett  $f = (r-1)(s-1)$  szabadsági fokú  $\chi^2$ -eloszláshoz tart:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(N_{ij} - \frac{N_{i\cdot} \cdot N_{\cdot j}}{n}\right)^2}{\frac{N_{i\cdot} \cdot N_{\cdot j}}{n}}.$$

Ha a függetlenség teljesül, akkor a  $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(B_j)$  egyenletben a valószínűségeket a relatív gyakoriságokkal helyettesítve:

$$\frac{N_{ij}}{n} \approx \frac{N_{i\cdot}}{n} \cdot \frac{N_{\cdot j}}{n} \quad \Leftrightarrow \quad N_{ij} \approx \frac{N_{i\cdot} \cdot N_{\cdot j}}{n}$$

Ebből adódik, hogy a  $\chi^2$  számlálója a nullhipotézistől való eltérést méri.

A szabadsági fok  $f = (r - 1)(s - 1)$ .

$c_{\text{krit}}$ : az  $f$  szabadsági fokú  $\chi^2$ -próba kritikus értéke  $\alpha$  szignifikanciaszint mellett.

- $\chi^2 < c_{\text{krit}}$  (azaz a  $p \geq \alpha$ ): elfogadjuk  $H_0$ -t, **nem találtunk szignifikáns összefüggést** a szempontok között.
- $\chi^2 > c_{\text{krit}}$  (azaz a  $p < \alpha$ ): elutasítjuk  $H_0$ -t, az adatok **szignifikáns összefüggést** mutatnak.

Ha  $r = s = 2$ , a próbastatisztika az alábbi egyszerűbb alakra hozható:

$$\chi^2 = \frac{n(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\cdot} \cdot N_{2\cdot} \cdot N_{\cdot 1} \cdot N_{\cdot 2}}$$

## 2.5. Függetlenségvizsgálat: példa

$H_0$ : a hőmérséklet és a csapadékmennyiség **független**;  $H_1$ : a hőmérséklet és a csapadékmennyiség között **összefüggés van**.

	meleg	átlagos	hideg
esős	15	10	5
átlagos	10	10	20
száraz	5	20	5

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(N_{ij} - \frac{N_{i\cdot} \cdot N_{\cdot j}}{n})^2}{\frac{N_{i\cdot} \cdot N_{\cdot j}}{n}} = \frac{(15 - \frac{30 \cdot 30}{100})^2}{\frac{30 \cdot 30}{100}} + \frac{(10 - \frac{30 \cdot 40}{100})^2}{\frac{30 \cdot 40}{100}} + \dots + \frac{(5 - \frac{30 \cdot 30}{100})^2}{\frac{30 \cdot 30}{100}} = 22,92$$

$n = 100$ ,  $f = (r - 1) \cdot (s - 1) = 2 \cdot 2 = 4$ ,  $\alpha = 0,05$ ,  $c_{\text{krit}} = 9,49$

$22,917 > c_{\text{krit}} = 9,49$ , illetve  $p = 0,00013 < \alpha = 0,05 \Rightarrow$  elutasítjuk a nullhipotézist, szignifikáns összefüggés van a két szempont között.

## 2.6. Pozitív korreláció

Tekintsük a függetlenségvizsgálatot abban az esetben, ha mindkét szempont szerint két osztály van. Ekkor az is értelmes kérdés, hogy „milyen irányú” az összefüggés, például igaz-e, hogy az  $A_1$  esemény a  $B_1$ -gyel pozitívan korrelál, azaz egyszerre nagyobb valószínűséggel következnek be, mint amit függetlenség esetén várnánk (ez utóbbi a két valószínűség szorzata lenne). Például, egy embert véletlenszerűen kiválasztva, van-e pozitív korreláció az alábbi két esemény között: van egyetemi végzettsége, a bruttó havi jövedelme legalább 400000 forint. Ez szorosan kapcsolódik a függetlenségvizsgálathoz, de nem  $\chi^2$ -próbát, hanem  $z$ -próbát tudunk alkalmazni.

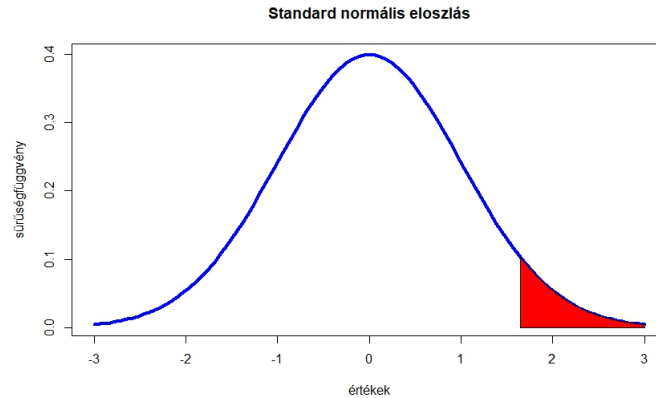
$H_0$ : a két szempont között **nincs pozitív korreláció**

$H_1$ : a két szempont között **pozitív korreláció** van, azaz  $\mathbb{P}(A_1 \cap B_1) > \mathbb{P}(A_1)\mathbb{P}(B_1)$ .

A próbastatisztika ( $H_0$  mellett standard normális eloszlású):

$$z = \sqrt{n} \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1\cdot} \cdot N_{2\cdot} \cdot N_{\cdot 1} \cdot N_{\cdot 2}}}$$

Ha  $z > \Phi^{-1}(1 - \alpha)$ , akkor elutasítjuk  $H_0$ -t, szignifikáns pozitív korreláció van; különben elfogadjuk  $H_0$ -t, nincs szignifikáns pozitív korreláció.



7. ábra. A standard normális eloszlás és a  $z$ -próba kritikus értéke  $\alpha = 0,05$  esetén

A  $p$ -érték:  $1 - \Phi(z)$ , ahol  $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$ . Vagyis a  $p$ -érték  $\int_z^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$ . Ez annak valószínűsége, hogy a standard normális eloszlás  $z$ -nél nagyobb, vagyis a 7. ábrán a  $z$ -től jobbra lévő terület.

### Pozitív korreláció: példa

Vérnyomás-szűrővizsgálatnál a 40 évesnél idősebbek közül 24-nek magas, 62-nek megfelelő volt a vérnyomása, a 40 évesnél nem idősebbek közül 12-nek volt magas, 88-nak megfelelő. Állíthatjuk-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy a 40 évesnél idősebbek között gyakoribb a magas vérnyomás?

$A_1$ : 40 évesnél nagyobb életkor;  $A_2$ : legfeljebb 40 éves életkor.

$B_1$ : magas vérnyomás;  $B_2$ : megfelelő vérnyomás.

$H_0$ : nincs pozitív korreláció;

$H_1$ : pozitív korreláció van.

$N_{11} = 24$ ;  $N_{12} = 62$ ;  $N_{21} = 12$ ;  $N_{22} = 88$ ;  $n = 186$ .

Minden osztályba esik legalább 5 megfigyelés ( $N_{ij} \geq 5$  minden  $i, j$ -re), alkalmazható a függetlenségvizsgálat.

$$z = \sqrt{n} \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}} = \sqrt{186} \cdot \frac{24 \cdot 88 - 62 \cdot 12}{\sqrt{86 \cdot 100 \cdot 36 \cdot 150}} = 2,74.$$

Mivel  $2,74 > \Phi^{-1}(0,95) = 1,645$ , így elutasítjuk a nullhipotézist. A nagyobb életkor és a magas vérnyomás között **szignifikáns pozitív** korreláció van. A  $p$ -érték:  $1 - \Phi(2,74) = 0,003 < 0,05$ .

A függetlenség vagy a pozitív korreláció vizsgálatánál a következőket érdemes figyelembe venni.

- minden osztályba essen legalább 5 megfigyelés
- a pozitív korreláció **nem jelent ok-okozati összefüggést**
- ha sok mennyiséget vizsgálunk, vagy előre el kell dönteni, hogy hol keressük a pozitív összefüggést: öt mennyiség között 10 pár van, így jó eséllyel lesz olyan pár, ahol tévesen szignifikáns összefüggést vagy pozitív korrelációt találhatunk ( $\alpha = 0,05$  szignifikanciaszintet választva); vagy alaposabban meg kell vizsgálni a kapott pozitív összefüggéseket (hiszen lehetnek tévesek); vagy a tévedés kockázatával számolva lehet használni a kapott összefüggéseket.

### 3. Homogenitásvizsgálat (kiegészítő anyag, a számonkérésnek nem része)

Legyenek  $X, Y$  valószínűségi változók,  $A_1, \dots, A_r$  teljes eseményrendszer.

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

$H_1$ : van legalább egy  $k$ , melyre  $\mathbb{P}(X \in A_k) \neq \mathbb{P}(Y \in A_k)$ .

$X_1, \dots, X_n, Y_1, \dots, Y_m$  független minta, melyre  $X_i \sim X, Y_i \sim Y$ .

$N_k$  az  $A_k$  gyakorisága az  $\underline{X}$  mintában;

$M_k$  az  $A_k$  gyakorisága az  $\underline{Y}$  mintában.

Ha  $N_k \geq 5$  vagy  $M_k \geq 5$  nem teljesül, osztályokat vonunk össze.

A próbastatisztika:

$$\chi^2 = \sum_{k=1}^r \frac{\left(\frac{N_k}{n} - \frac{M_k}{m}\right)^2}{\frac{N_k}{n} + \frac{M_k}{m}} \cdot n \cdot m.$$

Ha  $H_0$  igaz, és  $n \rightarrow \infty$ , akkor  $\chi^2$  eloszlása az  $f = r - 1$  szabadsági fokú  $\chi^2$ -eloszláshoz konvergál eloszlásban.

$c_{\text{krit}}$ : az  $f$  szabadsági fokú  $\chi^2$ -próba kritikus értéke  $\alpha$  terjedelem mellett.

- $\chi^2 < c_{\text{krit}}$  (azaz  $p \geq \alpha$ ): elfogadjuk  $H_0$ -t, nem találtunk szignifikáns eltérést az eloszlások között.
- $\chi^2 > c_{\text{krit}}$  (azaz a  $p < \alpha$ ): elutasítjuk  $H_0$ -t, az eloszlások szignifikánsan eltérnek.

#### 3.1. Homogenitásvizsgálat: példa

Két városban felmérték a háztartások létszámát, az elsőben  $n = 249$ , a másodikban  $m = 301$  elemű mintát vizsgálva. A szignifikanciaszintet  $\alpha = 0,05$ -nek választva állíthatjuk-e, hogy a két városban szignifikánsan eltérő a háztartások létszámának eloszlása?

létszám	1	2	3	4	$\geq 5$
első város	37	86	54	49	23
második város	45	94	67	56	39
első város, arány	0,15	0,35	0,22	0,2	0,09
második város, arány	0,18	0,38	0,27	0,22	0,16

Minden osztályba esik legalább 5 megfigyelés.

$$\chi^2 = \sum_{k=1}^r \frac{\left(\frac{N_k}{n} - \frac{M_k}{m}\right)^2}{\frac{N_k}{n} + \frac{M_k}{m}} \cdot n \cdot m = \left( \frac{(37/249 - 45/301)^2}{37 + 45} + \frac{(86/249 - 94/301)^2}{86 + 94} + \dots + \frac{(23/249 - 39/301)^2}{23 + 39} \right) \cdot 249 \cdot 301 = 2,23.$$

Az osztályok száma  $r = 5$ .

$$\chi^2 = 2,23; \quad f = r - 1 = 4; \quad \alpha = 0,05 \quad c_{\text{krit}} = 9,49$$

$\chi^2 = 2,23 < c_{\text{krit}} = 9,49$ , elfogadjuk a nullhipotézist, a két városban az egy háztartásban élők számának eloszlása **nem tér el szignifikánsan**. A  $p$ -érték:  $p = 0,31 > 0,05$ .

**Házi feladat április 15., 8:15-ig** A házi feladathoz begyűjtött adatsor alapján végezzük el az alábbi hipotézisvizsgálati feladatokat.

- Elfogadható-e  $\alpha = 0,05$  szignifikanciaszint mellett, hogy az egy ember által az elmúlt egy hónapban nézett sorozatok száma Poisson-eloszlású? (folytatás a következő oldalon)

- Számítsuk ki a nézett sorozatok számának mediánját a teljes adatsorból, legyen ez  $m$ . Állíthatjuk-e  $\alpha = 0,05$  valószínűséggel, hogy aközött, hogy egy véletlenszerűen kiválasztott ember nő, illetve hogy legalább  $m$  sorozatot nézett, szignifikáns összefüggés van?