

χ^2 -próba: illeszkedésvizsgálat (10. előadás)

Legyen A_1, A_2, \dots, A_r teljes eseményrendszer, 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.

- n független megfigyelést végzünk.
- N_k : hányszor következett be A_k .
- Ha van k , hogy $N_k < 4$: 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).
Ugyanakkor a próba túl nagy mintaelemszám túl érzékeny, kis eltérést is szignifikánsnak mutat.
- Próbastatisztika:

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

χ^2 -próba

Adott $(A_k)_{k=1}^r$ teljes eseményrendszer, és $(p_k)_{k=1}^r$ számok: $\sum_{k=1}^r p_k = 1$.

$H_0 : \mathbb{P}(A_k) = p_k$ minden $k = 1, 2, \dots, r$ -re. H_1 : a nullhipotézis nem igaz

Próbastatisztika (feltéve, hogy $N_k \geq 4$ minden k -ra):

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

χ^2 -próba

Adott $(A_k)_{k=1}^r$ teljes eseményrendszer, és $(p_k)_{k=1}^r$ számok: $\sum_{k=1}^r p_k = 1$.

$H_0 : \mathbb{P}(A_k) = p_k$ minden $k = 1, 2, \dots, r$ -re. H_1 : a nullhipotézis nem igaz

Próbastatisztika (feltéve, hogy $N_k \geq 4$ minden k -ra):

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

Legyen c_{krit} az $f = r - 1$ szabadsági fokú χ^2 -próba kritikus értéke α terjedelem (szignifikanciaszint) mellett.

$\chi^2 > c_{\text{krit}}$ 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.

χ^2 -próba

Adott $(A_k)_{k=1}^r$ teljes eseményrendszer, és $(p_k)_{k=1}^r$ számok: $\sum_{k=1}^r p_k = 1$.

$H_0 : \mathbb{P}(A_k) = p_k$ minden $k = 1, 2, \dots, r$ -re. H_1 : a nullhipotézis nem igaz

Próbastatisztika (feltéve, hogy $N_i \geq 4$ minden k -ra):

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

Legyen c_{krit} az $f = r - 1$ szabadsági fokú χ^2 -próba kritikus értéke α terjedelem (szignifikanciaszint) mellett.

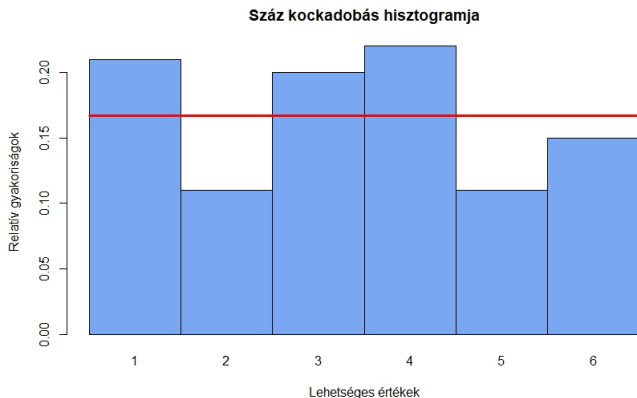
Ez az $f = r - 1$ szabadsági fokú χ^2 -eloszlás $1 - \alpha$ -kvantilise, vagyis

$$\mathbb{P}(Z_1^2 + \dots + Z_f^2 < c_{\text{krit}}) = 1 - \alpha,$$

ahol Z_1, \dots, Z_f független standard normális eloszlású valószínűségi változók.

χ^2 -próba: példa

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



χ^2 -próba: példa

Dobókockával dobunk százszor. 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

χ^2 -próba: példa

Dobókockával dobunk százszor. 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 χ^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

χ^2 -próba: példa

Dobókockával dobunk százszor. 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 χ^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

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

χ^2 -próba: példa

Dobókockával dobunk százszor. 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

χ^2 -próba: példa

Dobókockával dobunk százszor. 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

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

$$\chi^2 = 7,52; \quad f = r - 1 = 5; \quad \alpha = 0,05; \quad c_{\text{krit}} = 11,1$$

χ^2 -próba: példa

Dobókockával dobunk százszor. 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

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

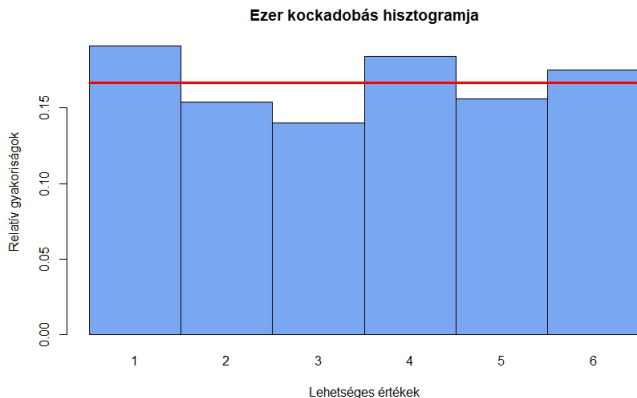
$$\chi^2 = 7,52; \quad f = r - 1 = 5; \quad \alpha = 0,05; \quad c_{\text{krit}} = 11,1$$

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

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

χ^2 -próba: példa

Dobókockával dobunk ezerszer. A terjedelmet $\alpha = 0,05$ -nek választva elfogadható-e, hogy szabályos a dobókocka?



χ^2 -próba: példa

Ha ezerszer dobunk, és az alábbi eredmények adódnak:

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

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

χ^2 -próba: példa

Ha ezerszer dobunk, és az alábbi eredmények adódnak:

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.

Pozitív korreláció

Igaz-e, hogy **két esemény gyakrabban fordul elő egyszerre**, mint ha függetlenek lennének egymástól?

Pozitív korreláció

Igaz-e, hogy **két esemény gyakrabban fordul elő egyszerre**, mint ha függetlenek lennének egymástól?

- minden osztályba essen legalább 5 megfigyelés
- a pozitív korreláció **nem jelent ok-okozati összefüggést**:
 - ▶ felhős ég, csapadék → pozitív korreláció és ok-okozati összefüggés is van;
 - ▶ télen felhőtlen égbolt, alacsony hőmérséklet →

Pozitív korreláció

Igaz-e, hogy **két esemény gyakrabban fordul elő egyszerre**, mint ha függetlenek lennének egymástól?

- minden osztályba essen legalább 5 megfigyelés
- a pozitív korreláció **nem jelent ok-okozati összefüggést**:
 - ▶ felhős ég, csapadék → pozitív korreláció és ok-okozati összefüggés is van;
 - ▶ télen felhőtlen égbolt, alacsony hőmérséklet → pozitív korreláció van, és ok-okozati összefüggés lehet mindkét irányban;
 - ▶ lazac rendszeres fogyasztása és egészség →

Pozitív korreláció

Igaz-e, hogy **két esemény gyakrabban fordul elő egyszerre**, mint ha függetlenek lennének egymástól?

- minden osztályba essen legalább 5 megfigyelés
- a pozitív korreláció **nem jelent ok-okozati összefüggést**:
 - ▶ felhős ég, csapadék → pozitív korreláció és ok-okozati összefüggés is van;
 - ▶ télen felhőtlen égbolt, alacsony hőmérséklet → pozitív korreláció van, és ok-okozati összefüggés lehet mindkét irányban;
 - ▶ lazac rendszeres fogyasztása és egészség → pozitív korreláció lehet, de ok-okozati összefüggés nincs feltétlenül, hanem mindkettő a jó anyagi helyzet következménye lehet
- ha sok eseménypár között végezzük a vizsgálatot, a öt mennyiség között (10 párnál) már jó eséllyel van tévesen szignifikáns pozitív korreláció (ha nincs összefüggés, akkor is mindegyik 5% valószínűséggel tévesen szignifikáns)
- big data analízis: sok eseménypár között végzik a vizsgálatot, lehetnek tévesen szignifikáns korrelációk, mégis használják ezeket

Pozitív korreláció

Két szempont szerint két-két osztályba soroljuk a megfigyeléseket. Tegyük fel, hogy mind a négy kategóriába esik legalább 5 megfigyelés.

H_0 : a két szempont között nincs pozitív korreláció

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

N_{ij} : az első szempont szerint i ., második szempont szerint j . osztályba eső megfigyelések száma. Továbbá $N_{i.} = N_{i1} + N_{i2}$, $N_{.j} = N_{1j} + N_{2j}$ és $n = N_{1.} + N_{2.}$.

A próbastatisztika (H_0 mellett standard normális eloszlású):

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

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

A p -érték: $1 - \Phi(u)$, ahol $\Phi(u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$.

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 $\alpha = 0,05$ terjedelem mellett, hogy a 40 évesnél idősebbek között gyakoribb a magas vérnyomás?

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 $\alpha = 0,05$ terjedelem 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.

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

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

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 $\alpha = 0,05$ terjedelem 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;

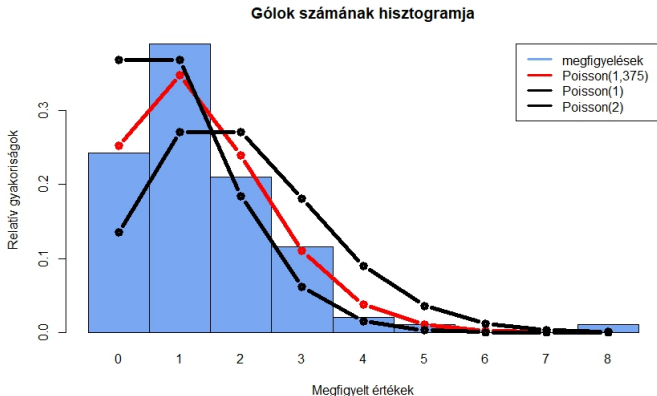
H_1 : pozitív korreláció van.

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

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

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: $1 - \Phi(2,74) = 0,003 < 0,05$.

Poisson-eloszlás paraméterének becslése



A gólok számának hisztogramja $n = 95$ mérkőzésen, és különböző paraméterű Poisson-eloszlások ($\mathbb{P}_\lambda(X = k) = \lambda^k / k! \cdot e^{-\lambda}$)

Becslések és tulajdonságaik

- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ valamely eloszlások halmaza valamilyen Θ halmazzal (Θ a paramétertér, a paraméter összes lehetséges értékeiből álló halmaz);
- $\psi : \Theta \rightarrow \mathbb{R}$ függvény.
- Cél: $\psi(\vartheta)$ becslése, azaz olyan T statisztika keresése, amire a $T(X_1, \dots, X_n)$ valószínűségi változó és a $\psi(\vartheta)$ érték valamilyen értelemben közel esnek egymáshoz.

Becslések és tulajdonságaik

- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ valamely eloszlások halmaza valamilyen Θ halmazzal (Θ a paraméterter, a paraméter összes lehetséges értékeiből álló halmaz);
- $\psi : \Theta \rightarrow \mathbb{R}$ függvény.
- Cél: $\psi(\vartheta)$ becslése, azaz olyan T statisztika keresése, amire a $T(X_1, \dots, X_n)$ valószínűségi változó és a $\psi(\vartheta)$ érték valamilyen értelemben közel esnek egymáshoz.

Definíció (Torzítatlanság)

A T statisztika torzítatlan becslés ψ -re, ha minden $\vartheta \in \Theta$ -ra

$$\mathbb{E}_\vartheta(T(X_1, \dots, X_n)) = \psi(\vartheta).$$

A T statisztika torzítása a $b_T(\vartheta) = \mathbb{E}_\vartheta(T(X_1, \dots, X_n)) - \psi(\vartheta)$ függvény.

Példa. X_1, X_2, \dots, X_n független minta a $[0, \vartheta]$ intervallumon egyenletes eloszlásból. Ekkor $2\bar{X}$ torzítatlan becslés $\psi(\vartheta) = \vartheta$ -ra.

Torzítatlan becslések

Állítás (A várható érték torzítatlan becslése)

Legyen X_1, \dots, X_n független azonos eloszlású véges várható értékű minta. Ekkor

$$\mathbb{E}_\vartheta(\bar{X}) = \mathbb{E}_\vartheta(X_1) \quad \text{minden } \vartheta \in \Theta\text{-ra,}$$

vagyis a **mintaátlag** torzítatlan becslés ψ -re.

Állítás (A szórásnégyzet torzítatlan becslése)

X_1, \dots, X_n független azonos eloszlású véges szórású minta. Ekkor Ekkor

$$\mathbb{E}_\vartheta(s_n^{*2}) = D_\vartheta^2(X_1) \quad \text{minden } \vartheta \in \Theta\text{-ra,}$$

vagyis a **korrigált tapasztalati szórásnégyzet** torzítatlan becslés a szórásnégyzet-re.

Becslések összehasonlítása

Definíció (Hatásosság)

Legyenek T_1, T_2 **torzítatlan** becslései a paraméter $\psi(\vartheta)$ függvényének. T_1 **hatásosabb** T_2 -nél, ha

$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

teljesül minden $\vartheta \in \Theta$ -ra.

A T_1 becslés **hatásos** $\psi(\vartheta)$ -ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).

Becslések összehasonlítása

Definíció (Hatásosság)

Legyenek T_1, T_2 **torzítatlan** becslései a paraméter $\psi(\vartheta)$ függvényének. T_1 **hatásosabb** T_2 -nél, ha

$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

teljesül minden $\vartheta \in \Theta$ -ra.

A T_1 becslés **hatásos** $\psi(\vartheta)$ -ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).

- Nem mindig létezik hatásos becslés, és lehetséges, hogy T_1 és T_2 közül egyik sem hatásosabb a másiknál.
- A várható értékre nézve a mintaátlag hatásosabb minden $\sum_{j=1}^n c_j X_j$ alakú becslésnél (ahol $\sum_{j=1}^n c_j = 1$).

Becslések összehasonlítása

Definíció (Hatásosság)

Legyenek T_1, T_2 **torzítatlan** becslései a paraméter $\psi(\vartheta)$ függvényének. T_1 **hatásosabb** T_2 -nél, ha

$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

teljesül minden $\vartheta \in \Theta$ -ra.

A T_1 becslés **hatásos** $\psi(\vartheta)$ -ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).

- Nem mindig létezik hatásos becslés, és lehetséges, hogy T_1 és T_2 közül egyik sem hatásosabb a másiknál.
- A várható értékre nézve a mintaátlag hatásosabb minden $\sum_{j=1}^n c_j X_j$ alakú becslésnél (ahol $\sum_{j=1}^n c_j = 1$).
- **Bizonyos feladatokban lehet a mintaátlagnál hatásosabb becslés a várható értékre:** A $[0, b]$ intervallumon egyenletes eloszlás esetén b -re $\frac{n+1}{n} \max(X_1, \dots, X_n)$ hatásosabb a mintaátlag kétszeresénél.