

χ^2 -próbák (9. előadás)

- illeszkedésvizsgálat: diszkrét eloszlások illeszkedésének ellenőrzése, egy adott eloszlásból származnak-e?
- függetlenségvizsgálat: a megfigyeléseket két szempont szerint véges sok kategóriába soroljuk be – független-e a két szempont?
- homogenitásvizsgálat diszkrét eloszlásokra: azonos-e két minta eloszlása?

Tulajdonságok

- a χ^2 -próba **aszimptotikus próba**, vagyis a próbastatisztika eloszlása nem pontosan χ^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 mintaelemszámnál** a χ^2 -próba **túlságosan érzékeny**, túl gyakran mutat ki szignifikáns eltérést

χ^2 -próba: illeszkedésvizsgálat

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).
- 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 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):

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

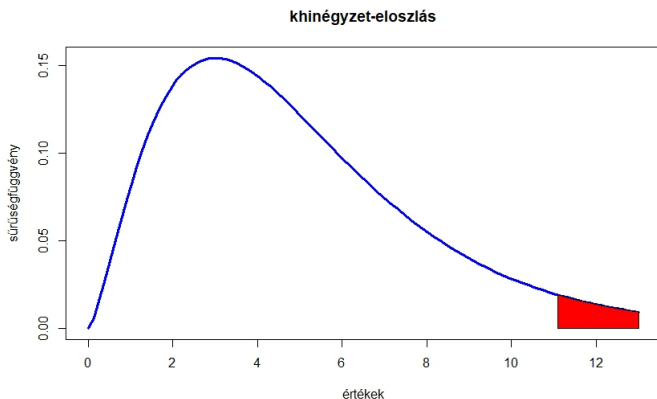
Legyen c_{krit} az $f = r - 1$ szabadsági fokú χ^2 -próba kritikus értéke α terjedelem (szignifikanciaszint) mellett.

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.

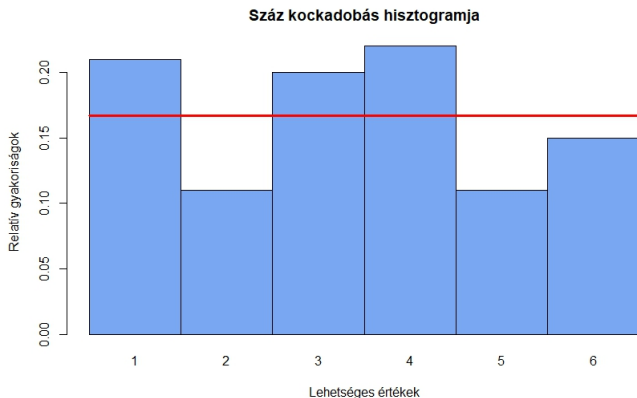
χ^2 -próba kritikus értéke



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

χ^2 -próba: példa

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



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

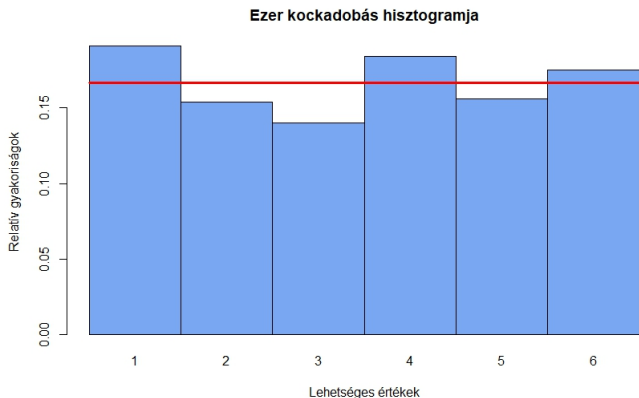
$$\chi^2 = 7,52; \quad f = r - 1 = 5; \quad \alpha = 0,05; \quad c_{\text{krit}} = 11,1$$

$\chi^2 = 7,52 < c_{\text{krit}} = 11,1$, 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

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

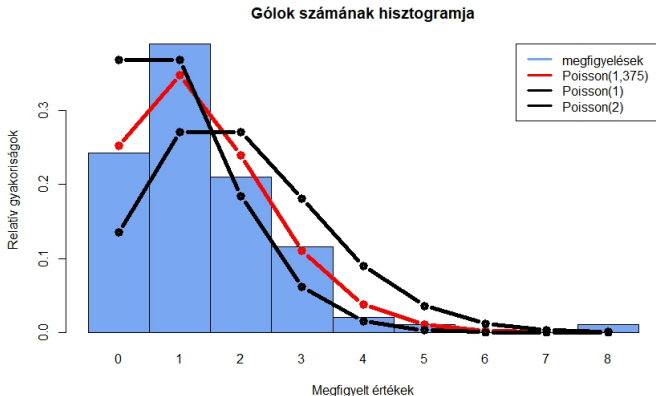
$\chi^2 = 11,68 > c_{\text{krit}} = 11,1$, illetve a p -értékre $0,039 < 0,05$.

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

Becsléses illeszkedésvizsgálat: példa

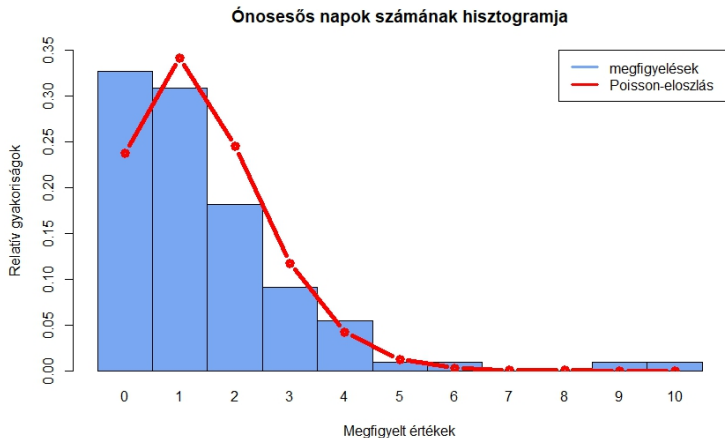
Elfogadható-e 0,05 terjedelem (szignifikanciaszint) mellett, hogy az egy futballmérkőzésen lőtt gólok száma Poisson-eloszlású?

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

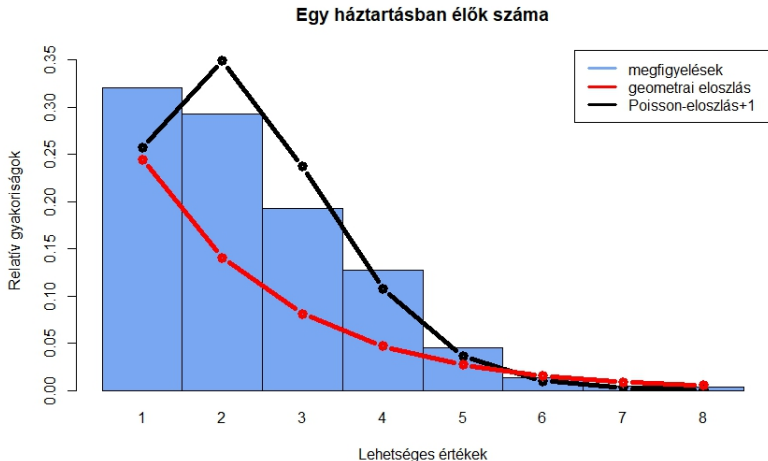


Becsléses illeszkedésvizsgálat: példa

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



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.**

Becsléses illeszkedésvizsgálat

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 $N_k \geq 4$ 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.

Becsléses illeszkedésvizsgálat

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 $N_k \geq 4$ 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 λ paramétervektor maximumlikelihood-becslése legyen $\hat{\lambda}$, és legyen $\hat{p}_k = p_k(\hat{\lambda})$.
A λ 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_{krit} az f szabadsági fokú χ^2 -próba kritikus értéke α 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.

Becsléses illeszkedésvizsgálat: példa

Példa. Az egy futballmérkőzésen lőtt gólok száma a világbajnokság $n = 95$ mérkőzésén:

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 λ 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 4-nél kevesebb megfigyelés esik, a beosztást módosítjuk:

gólok száma	0	1	2	3	≥ 4
mérkőzések száma	23	37	20	11	4

Becsléses illeszkedésvizsgálat: példa

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.

Becsléses illeszkedésvizsgálat: példa

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	≥ 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

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

Becsléses illeszkedésvizsgálat: példa

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.

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

Becsléses illeszkedésvizsgálat: példa

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

H_1 : az eloszlás eltér a Poisson-eloszlástól.

$\hat{\lambda} = 1,379$, egydimenziós paramétert (egy pozitív számot) kellett becsülni, tehát $d=1$. Az osztályok száma $r = 5$.

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

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

Becsléses illeszkedésvizsgálat: példa

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

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

$\hat{\lambda} = 1,379$, egydimenziós paramétert (egy pozitív számot) kellett becsülni, tehát $d=1$. Az osztályok száma $r = 5$.

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

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

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

Becsléses illeszkedésvizsgálat: példa

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 λ 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 4-nél kevesebb megfigyelés esik, a beosztást módosítjuk:

ónosesős napok száma	0	1	2	3	4	≥ 5
évek száma	36	34	20	10	6	4

Becsléses illeszkedésvizsgálat: példa

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	≥ 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

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

Becsléses illeszkedésvizsgálat: példa

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

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

$\hat{\lambda} = 1,436$, egydimenziós paramétert (egy pozitív számot) kellett becsülni, tehát $d=1$. Az osztályok száma $r = 6$.

ónosesős napok száma	0	1	2	3	4	≥ 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

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

Becsléses illeszkedésvizsgálat: példa

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

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

$\hat{\lambda} = 1,436$, egydimenziós paramétert (egy pozitív számot) kellett becsülni, tehát $d=1$. Az osztályok száma $r = 6$.

ónosesős napok száma	0	1	2	3	4	≥ 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

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

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

Függetlenségvizsgálat

Két szempont szerint soroljuk osztályokba a megfigyeléseket.

Első szempont: A_1, \dots, A_r . Második szempont: B_1, \dots, B_s .

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.

Függetlenségvizsgálat

Két szempont szerint soroljuk osztályokba a megfigyeléseket.

Első szempont: A_1, \dots, A_r . Második szempont: B_1, \dots, B_s .

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.

$N_{i.} = \sum_{j=1}^s N_{ij}$ (azaz az A_i gyakorisága); $N_{.j} = \sum_{i=1}^r N_{ij}$ (azaz B_j gyakorisága); n pedig az összes megfigyelés száma. Ekkor a próbat statisztika:

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

Függetlenségvizsgálat

H_0 : a két szempont független egymástól. Próbastatisztika:

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

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

c_{krit} : az f szabadsági fokú χ^2 -próba kritikus értéke α 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.

Függetlenségvizsgálat

H_0 : a két szempont független egymástól. Próbastatisztika:

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

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

c_{krit} : az f szabadsági fokú χ^2 -próba kritikus értéke α 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:

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

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

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

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

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

Pozitív korreláció

Tekintsük a függetlenségvizsgálatot abban az esetben, ha mindkét szempont szerint két osztály van.

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

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

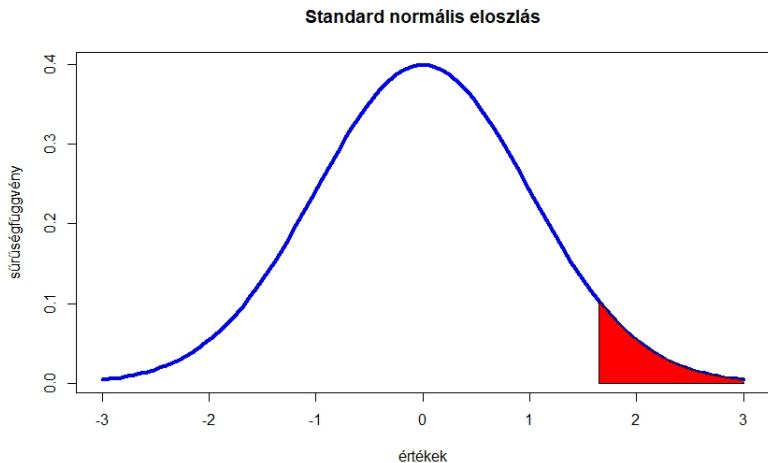
A próbastatisztika (H_0 mellett standard normális eloszlású):

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

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

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

Az egyoldali z-próba kritikus értéke



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

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?

Pozitív korreláció: példa

Vérnyomás-szűróvizsgálatnál 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.

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

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

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.

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

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

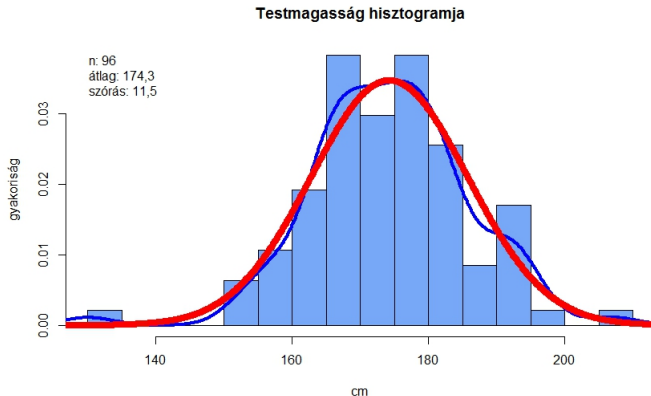
Mivel $2,74 > \Phi^{-1}(0,95) = 1,645$, így elutasítjuk a nullhipotézist. A nagyobb életkor és a magas vérnyomás között **szignifikáns pozitív** korreláció van. A p -érték: $1 - \Phi(2,74) = 0,003 < 0,05$.

Pozitív korreláció

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

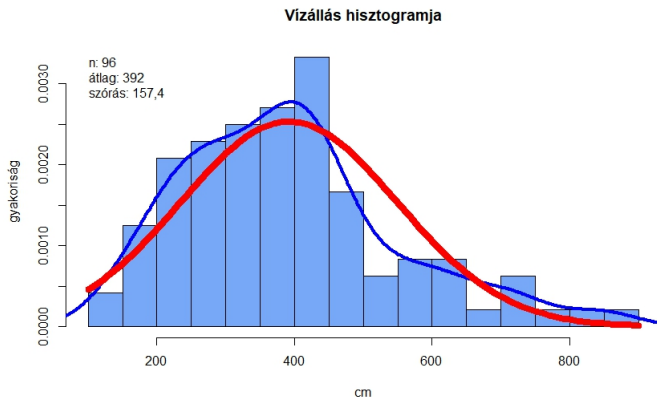
- minden osztályba essen legalább 6 megfigyelés
- a pozitív korreláció **nem jelent ok-okozati összefüggést**
- ha sok mennyiséget vizsgálunk, előre kell eldönteni (az adatok ismerete nélkül), hogy hol keressük a pozitív összefüggést: öt mennyiség között 10 pár van, így jó eséllyel lesz olyan pár, ahol tévesen szignifikáns összefüggést vagy pozitív korrelációt találhatunk ($\alpha = 0,05$ szignifikanciaszintet választva)

Testmagasság és normális eloszlás



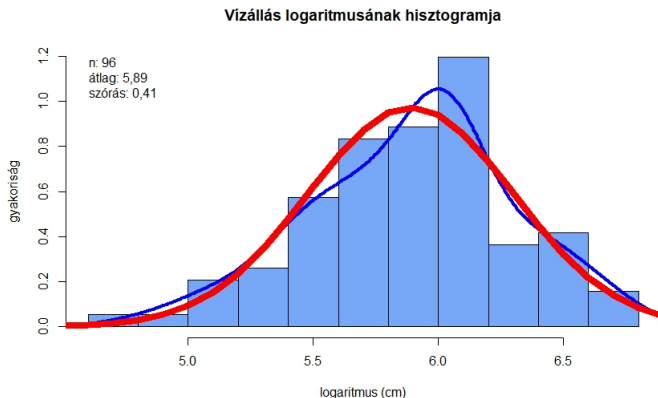
A testmagasság hisztogramja $n = 96$ elemű mintából, a sűrűségfüggvény becslése Gauss-magfüggvénnyel, és az $\bar{X} = 174,3$ várható értékű és $s_n^* = 11,5$ szórású normális eloszlás sűrűségfüggvénye.

A Duna vízállása



A Duna havi legnagyobb vízállásának hisztogramja (2002–2009, $n = 96$, forrás: Országos Vízelző Szolgálat), a becsült sűrűségfüggvény, és az $\bar{X} = 392$ várható értékű és $s_n^* = 157,4$ szórású normális eloszlás sűrűségfüggvénye – **itt a függetlenség nem teljesen érvényes**

A Duna vízállása



A Duna havi legnagyobb vízállásának **logaritmusának** hisztogramja (2002–2009, $n = 96$, forrás: Országos Vízeljáró Szolgálat), a becsült sűrűségfüggvény, és az $\bar{X} = 392$ várható értékű és $s_n^* = 157,4$ szórású normális eloszlás sűrűségfüggvénye

Nemparaméteres próbák

Illeszkedésvizsgálat: a minta egy adott, folytonos eloszlásból származik-e?

Homogenitásvizsgálat: két minta ugyanabból az eloszlásból származik-e?

Egy lehetőség: **diszkrétizáljuk** a megfigyeléseket, vagyis közel azonos hosszúságú intervallumokba osztjuk be őket, és az így kapott diszkrét eloszlásra χ^2 -próbát végzünk. Ha szükséges, a paramétereket maximumlikelihood-módszerrel becsüljük.

Tapasztalati eloszlásfüggvények távolságát használó próbák:

- Kolmogorov–Szmirnov-próba
- Anderson–Darling-próba (más súlyozás)
- Cramér–von Mises próba (más súlyozás)

Speciálisan annak ellenőrzésére, hogy egy eloszlás **normális eloszlású**-e:

- Lilliefors-próba (a Kolmogorov–Szmirnov-próbán alapul)
- Shapiro–Wilk-próba (a rendezett minta várható értékét és kovarianciamátrixát használja)