

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?

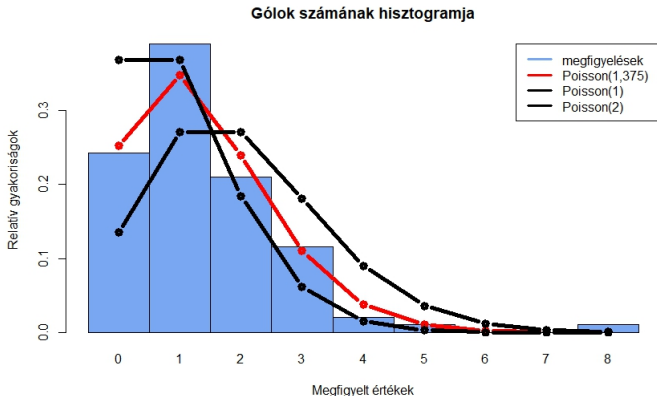
Tulajdonságok

- 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

Becsléses illeszkedésvizsgálat: példa

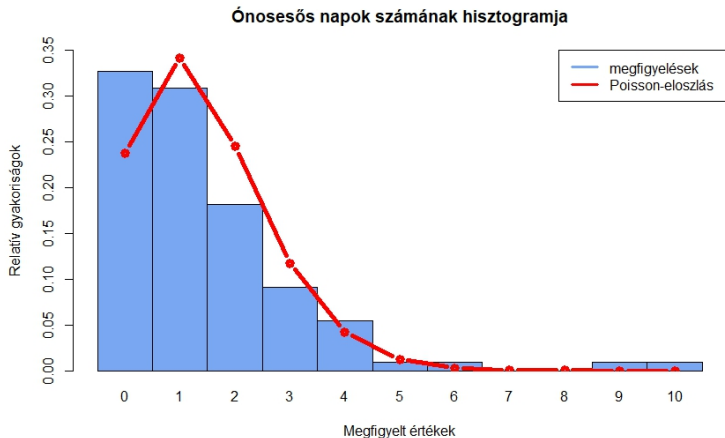
Elfogadható-e 0,05 terjedelem (szignifikanciaszint) mellett, hogy az egy futballmérkőzésen lőtt gólok száma Poisson-eloszlású?

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

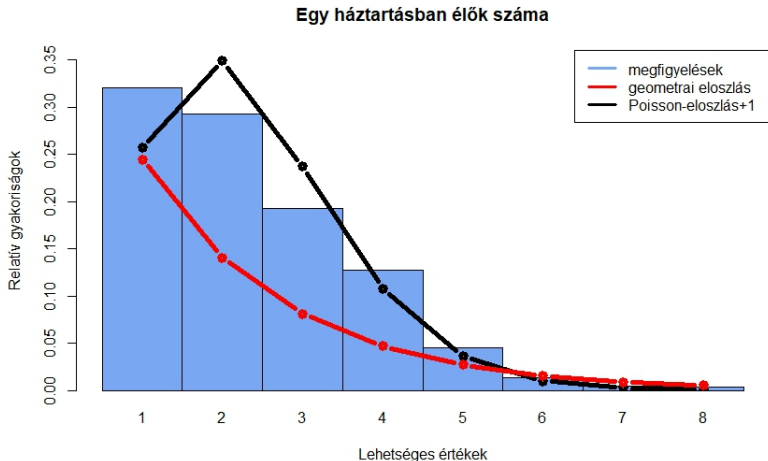


Becsléses illeszkedésvizsgálat: példa

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

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

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

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 $\mathbb{P}(A_1 \cap B_1) > \mathbb{P}(A_1)\mathbb{P}(B_1)$.

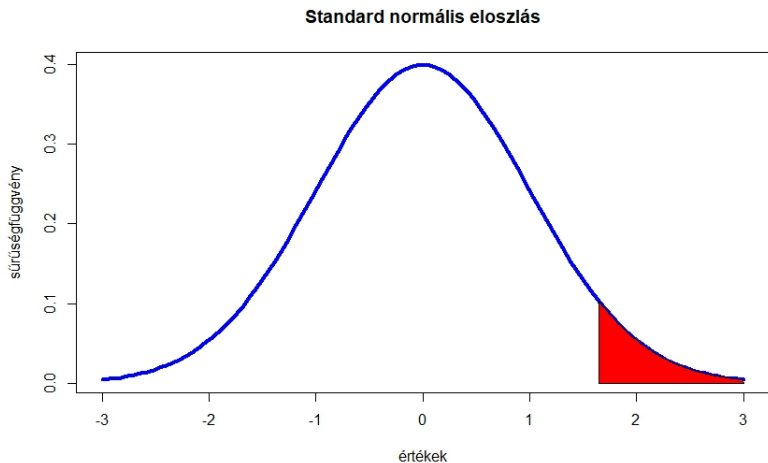
A próbastatisztika (H_0 mellett standard normális eloszlású):

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

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)

χ^2 -próba: homogenitásvizsgálat

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 $N_k \geq 4$ vagy $M_k \geq 4$ nem teljesül, osztályokat vonunk össze.

A próbastatisztika:

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

Homogenitásvizsgálat

A próbastatisztika:

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

A szabadsági fok: $f = r - 1$.

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

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	> 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

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	> 4
első város	37	86	54	49	23
második város	45	94	67	56	39

Minden osztályba esik legalább 4 megfigyelés.

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

Homogenitásvizsgálat: példa

Két városban felmérték a háztartások létszámát. A szignifikanciaszintet $\alpha = 0,05$ -nek választva állíthatjuk-e, hogy a két városban szignifikánsan eltérő a háztartások létszámának eloszlása?

létszám	1	2	3	4	> 4
első város	37	86	54	49	23
második város	45	94	67	56	39

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 kétféle homok szemcseméretének eloszlása **nem tér el szignifikánsan**. A p -érték: $p = 0,31 > 0,05$.

Házi feladat április 24., 9:00-ig

- 1 A húszelemű minta alapján állíthatjuk-e $\alpha = 0,05$ szignifikanciaszint mellett, hogy szignifikáns pozitív korreláció van aközött, hogy valakinek legalább 175 cm a testmagassága, és legalább 40-es a cipőmérete? Határozzuk meg a p -értéket is.

A 175 és 40 helyett választhatunk más, tetszőleges határokat, viszont, ha lehetséges, minden osztályba essen legalább 3 megfigyelés (a kevés adat miatt a legalább 6 most nem lenne megvalósítható).

- 2 A húszelemű minta alapján állíthatjuk-e $\alpha = 0,05$ szignifikanciaszint mellett, hogy a nem és a cipőméret szignifikánsan összefüggő szempontok?

Házi feladat április 10., 9:00-ig

Tekintsünk egy tetszőleges valószínűségeloszlást, azaz hat pozitív számot, aminek az összege 1. Legyenek ezek p_1, p_2, \dots, p_6 , és egyik se legyen $1/6$ -dal egyenlő, és egyik se legyen $1/10$ -nél kisebb.

- Tekintsünk 100000 független kockadobást, ahol az i szám valószínűsége a fent választott p_i . Csoportosítsuk százasával a mintaelemeket, és minden százas csoportban számítsuk ki a χ^2 -statisztika értékét:

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

Készítsünk hisztogramot az így kapott ezer értékből, és hasonlítsuk ezt össze az $f = 5$ szabadsági fokú χ^2 -eloszlás sűrűségfüggvényével.

- Tekintsünk 5000 független standard normális eloszlású valószínűségi változót. Csoportosítsuk ötösével a mintaelemeket, és minden ötös csoportban számítsuk ki az elemek négyzetének összegét. Készítsünk hisztogramot az így kapott ezer értékből, és hasonlítsuk ezt össze az $f = 5$ szabadsági fokú χ^2 -eloszlás sűrűségfüggvényével.

Házi feladat április 10-ig: megoldás

```
p=c(0.2, 0.1, 0.2, 0.25, 0.15, 0.1)
```

```
x<-sample(1:6, 100000, prob=p, replace=T)
```

```
m<-matrix(x, ncol=100)
```

```
ossz<-function(j)
```

```
{v=1:6; for (k in 1:6) {v[k]=sum(m[j,]==k)}; sum((v-100*p)^2/100/p)}
```

```
eredmeny<-sapply(1:1000, ossz)
```

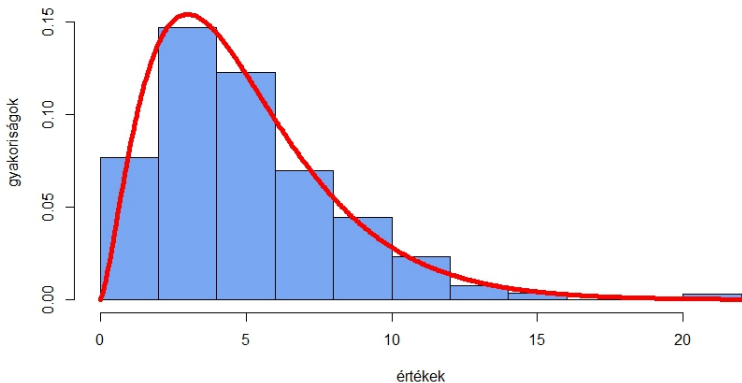
```
hist(eredmeny, col="#79a7f2", xlab="értékek", ylab="gyakoriságok", main="Khí-  
négyzet próba statisztikájának eloszlása", freq=F, ylim=c(0, 0.17))
```

```
x=seq(from=0, to=22, by=0.05); y=dchisq(x, df=5);
```

```
lines(y~ x, col="red", lwd="5")
```

Házi feladat április 10-ig: megoldás

Khi-négyzet próba statisztikájának eloszlása



A χ^2 -statisztikák eloszlása az $(0, 2; 0, 1; 0, 2; 0, 25; 0, 15; 0, 1)$ valószínűségeloszlásból kiindulva, és az $f = 5$ szabadsági fokú χ^2 -eloszlás sűrűségfüggvénye

Házi feladat április 10-ig: megoldás

```
x<-rnorm(5000)
```

```
m<-matrix(x, ncol=5)
```

```
ossz<-function(j) {sum(m[,j]*m[,j])}
```

```
eredmeny<-sapply(1:1000, ossz)
```

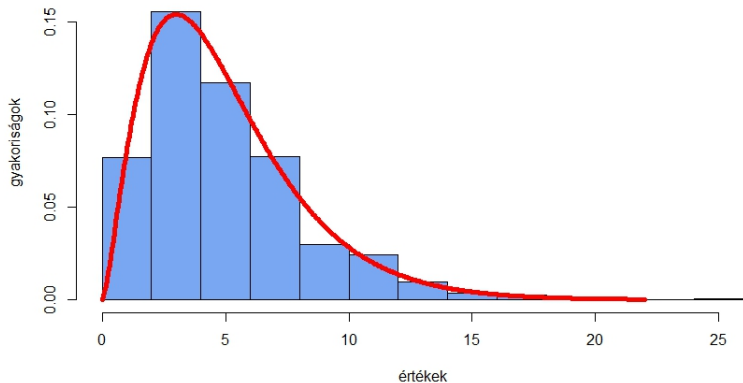
```
hist(eredmeny, col="#79a7f2", xlab="értékek", ylab="gyakoriságok", main="Öt  
standard normális eloszlás négyzetösszege", freq=F, ylim=c(0, 0.17))
```

```
x=seq(from=0, to=22, by=0.05); y=dchisq(x, df=5);
```

```
lines(y~ x, col="red", lwd="5")
```

Házi feladat április 10-ig: megoldás

Öt standard normális eloszlás négyzetösszege



Öt darab független standard normális eloszlás négyzetösszegének eloszlása és az $f = 5$ szabadsági fokú χ^2 -eloszlás sűrűségfüggvénye