7. feladatsor, valószínűségszámítás, 2019. november 13.

1. A születések száma 2016-ban az alábbiak szerint alakult Magyarországon (forrás: Központi Statisztikai Hivatal).

évszak	tavasz	nyár	ősz	tél
születések száma	21337	24299	24691	22736

Ezek alapján $0,01$ terjedelem (szignifikanciaszint) mellett elfogadható-e az a hipotézis, hogy a születések időpontja minden évszakba azonos, $1/4$ valószínűséggel esik?

Illeszkedésvizsgálatot végzünk χ^2-próbával. A megfigyelések függetlenek, minden osztályba esik legalább 4 megfigyelés, nem kell osztályokat összevonni. Az osztályok száma $r = 4$, és $p_1 = p_2 = p_3 = p_4 = 1/4$, mert minden évszakról az $1/4$ valószínűséget tételezzük fel.

H_0: minden évszak valószínűsége $p_k = 1/4$

H_1: van olyan évszak, aminek nem $1/4$ a valószínűsége

A megfigyelések száma $n = 93063$. A próbastatisztika:

$$\chi^2 = \sum_{k=1}^{r} \frac{(N_k - np_k)^2}{np_k} = \frac{(21337 - 93063 \cdot 1/4)^2}{93063 \cdot 1/4} + \frac{(24299 - 93063 \cdot 1/4)^2}{93063 \cdot 1/4} + \frac{(24961 - 93063 \cdot 1/4)^2}{93063 \cdot 1/4} + \frac{(22736 - 93063 \cdot 1/4)^2}{93063 \cdot 1/4} = 305,15.$$

A próba szabadsági foka: $f = r - 1 = 3$, a terjedelme $\alpha = 0,01$, így a kritikus érték: $c_{\text{krit}} = 11,3$.

Mivel $\chi^2 = 305,15 > 11,3 = c_{\text{krit}}$, elutasítjuk a nullhipotézist, a születések száma az egyes évszakokban szignifikánsan eltér az egyenletes eloszlástól. (A p-érték kisebb 10^{-15}-nél.)

Ez a példa azt is mutatja, hogy a χ^2-próba nagy mintaelemszám esetén nagyon érzékeny a kis eltérésekre, valójában ilyenkor nem is mindig alkalmazható.

2. Tegyük fel, hogy egy bizonyos helyszínen a homokban az 1 mm-nél nagyobb szemcsék aránya 22%, a $0,5 - 1$ mm nagyságúaké 34%, a $0,25 - 0,5$ mm nagyságúaké 26%, a többi 18% pedig $0,25$ mm-nél is kisebb. Van egy minta, amiről el szeretnénk dönteni, hogy származhat-e erről a helyszínről, vagy a szemcseméretének eloszlása szignifikánsan eltér az itt megadottól. A mintában 300 szemcse méretét határozták meg, ezek közül 72 darab volt 1 mm-nél nagyobb, 119 darab volt $0,5 - 1$ mm között, 78 darab volt $0,25 - 0,5$ közötti méretű. Ez alapján állíthatjuk-e $0,05$ terjedelem (szignifikanciaszint) mellett, hogy a minta szemcseméretének eloszlása szignifikánsan eltér a megadott helyszínen tapasztalható eloszlástól?

Illeszkedésvizsgálatot végzünk χ^2-próbával. A megfigyelések függetlenek, minden osztályba esik legalább 4 megfigyelés, nem kell osztályokat összevonni. Az osztályok száma $r = 4$, és $p_1 = 0,22$, $p_2 = 0,34$, $p_3 = 0,26$, $p_4 = 0,18$.

H_0: a k. kategória valószínűsége p_k minden $k = 1, 2, 3, 4$-re

H_1: van olyan kategória, aminek a valószínűsége nem a megadott valószínűség

A megfigyelések száma $n = 300$. A próbastatisztika:

$$\chi^2 = \sum_{k=1}^{r} \frac{(N_k - np_k)^2}{np_k} = \frac{(72 - 300 \cdot 0,22)^2}{300 \cdot 0,22} + \frac{(119 - 300 \cdot 0,34)^2}{300 \cdot 0,34} + \frac{(78 - 300 \cdot 0,26)^2}{300 \cdot 0,26} + \frac{(31 - 300 \cdot 0,18)^2}{300 \cdot 0,18} = 13,175.$$

A próba szabadsági foka: $f = r - 1 = 3$, a terjedelme $\alpha = 0,05$, így a kritikus érték: $c_{\text{krit}} = 7,81$.

Mivel $\chi^2 = 13,175 > 7,81 = c_{\text{krit}}$, elutasítjuk a nullhipotézist, a minta szemcsemérete szignifikánsan eltér a megadott eloszlástól. (A p-érték $0,004 < 0,05 = \alpha$.)

3. Néhány (azonos méretű) kőzetmintában megvizsgálták a fosszíliákat. Az alábbi táblázat azt tartalmazza, hogy hány olyan kőzetminta volt, melyben adott számú élőlény nyomát találták meg.

élőlények száma	0	1	2	3	4	5	6
kőzetek száma	2	5	7	5	7	3	1

$\alpha = 0,05$ terjedelem (szignifikanciaszint) mellett elfogadható-e az a hipotézis, hogy az egy kőzetmintában található fosszíliák száma Poisson-eloszlású?

Becsléses illeszkedésvizsgálatot végzünk χ^2-próbával.

H_0: a fosszíliák száma egy kőzetmintában Poisson-eloszlású

H_1: a fosszíliák számának eloszlása eltér a Poisson-eloszlástól

A Poisson-eloszlás definíciója: $p_k = \frac{\lambda^k}{k!} e^{-\lambda}$, ahol $k = 0, 1, 2 \ldots$. Ebben egy ismeretlen paraméter van, tehát $s = 1$ lesz, és λ-t kell maximumlikelihood-módszerrel becsülni, vagyis az átlagot kiszámítani (összesen $n = 30$ megfigyelés volt):

$$\hat{\lambda} = \frac{0 \cdot 2 + 1 \cdot 5 + 2 \cdot 7 + 3 \cdot 5 + 4 \cdot 7 + 5 \cdot 3 + 6 \cdot 1}{30} = 2,77.$$

Nincs minden osztályban legalább 4 megfigyelés, ezért a χ^2-próba elvégzése előtt össze kell vonni osztályokat, hogy legalább 4 legyen mindenhol, majd kiszámolni a $\hat{p}_k$-t, ami $\frac{\hat{\lambda}^k}{k!}e^{-\hat{\lambda}}$-val egyenlő, illetve a $0-1$ esetben $e^{-\hat{\lambda}}+\hat{\lambda}e^{-\hat{\lambda}}$ (a $k=0$ és $k=1$ összege), a ≥ 5 valószínűséget pedig úgy kapjuk, hogy egy legyen összeg:

élőlények száma (k)	$0-1$	2	3	4	≥ 5
kőzetek száma (N_k)	7	7	5	7	4
$\hat{p}_k$	0,24	0,24	0,22	0,15	0,15
np_k	7,2	7,2	6,6	4,5	4,5

Tehát

$$\chi^2=\sum_{k=1}^{r}\frac{(N_k-np_k)^2}{np_k}=\frac{(7-7,2)^2}{7,2}+\frac{(7-7,2)^2}{7,2}+\frac{(5-5,6)^2}{5,6}+\frac{(7-4,5)^2}{4,5}+\frac{(4-4,5)^2}{4,5}=1,52.$$

Az osztályok száma: $r=5$ (az összevonás utáni érték számít). Az $f=r-s-1=5-1-1=3$ szabadsági fokú (egy paramétert becsültünk, ezért $s=1$), $\alpha=0,05$ terjedelmű χ^2-próba kritikus értéke: $c_{\text{krit}}=7,81$.

Tehát $\chi^2=1,52<c_{\text{krit}}=7,81$, elfogadható az a hipotézis, hogy a fosszíliák száma Poisson-eloszlású.

4. Az alábbi táblázatok azt mutatják, hogy milyen volt egy gyógyszer használatának eredménye, amikor férfiakon, illetve nőkön próbálták ki.

férfiak	gyógyszer	placebo
gyógyult	700	80
nem gyógyult	800	130

nők	gyógyszer	placebo
gyógyult	150	400
nem gyógyult	70	280

Csak a férfiakat figyelembe véve, állíthatjuk-e $\alpha=0,01$ terjedelem (szignifikanciaszint) mellett, hogy a gyógyszeres kezelés és a gyógyulás között szignifikáns pozitív korreláció van? Mi a válasz, ha csak a nőket vesszük figyelembe?

Ha összesítjük az adatokat, állíthatjuk-e $\alpha=0,01$ terjedelem mellett, hogy a gyógyszeres kezelés és a gyógyulás között szignifikáns pozitív korreláció van?

Minden osztályba esik legalább négy megfigyelés, alkalmazható a pozitív korrelációra vonatkozó módszer.

H_0: $\mathbb{P}$(gyógyszeres kezelés és gyógyulás) $\leq \mathbb{P}$(gyógyszeres kezelés) $\cdot\, \mathbb{P}$(gyógyulás)

H_1: $\mathbb{P}$(gyógyszeres kezelés és gyógyulás) $> \mathbb{P}$(gyógyszeres kezelés) $\cdot\, \mathbb{P}$(gyógyulás)

A férfiak esetében:

$$u=\sqrt{n}\frac{N_{11}N_{22}-N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}}=\sqrt{1710}\cdot\frac{700\cdot 130-800\cdot 80}{\sqrt{780\cdot 930\cdot 1500\cdot 210}}=2,336>\Phi^{-1}(1-\alpha)=\Phi^{-1}(0,99)=2,326.$$

ahol a kritikus érték a standard normális eloszlásfüggvényének inverze az $1-\alpha$ helyen. Tehát elutasítjuk a nullhipotézist, szignifikáns pozitív korreláció van a gyógyszeres kezelés és a gyógyulás között a férfiak esetében.

A nők esetében:

$$u=\sqrt{n}\frac{N_{11}N_{22}-N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}}=\sqrt{900}\cdot\frac{150\cdot 280-400\cdot 70}{\sqrt{550\cdot 350\cdot 220\cdot 680}}=2,47>\Phi^{-1}(1-\alpha)=\Phi^{-1}(0,99)=2,326.$$

ahol a kritikus érték a standard normális eloszlásfüggvényének inverze az $1-\alpha$ helyen. Tehát elutasítjuk a nullhipotézist, szignifikáns pozitív korreláció van a gyógyszeres kezelés és a gyógyulás között a nők esetében is.

Összesítve az adatokat:

összesen	gyógyszer	placebo
gyógyult	850	480
nem gyógyult	870	410

Az összesített adatokból:

$$u=\sqrt{n}\frac{N_{11}N_{22}-N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}}=\sqrt{2610}\cdot\frac{850\cdot 410-870\cdot 480}{\sqrt{1330\cdot 1280\cdot 1720\cdot 890}}=-2,19\leq\Phi^{-1}(1-\alpha)=\Phi^{-1}(0,99)=2,326.$$

ahol a kritikus érték a standard normális eloszlásfüggvényének inverze az $1-\alpha$ helyen. Tehát elfogadjuk a nullhipotézist, nincs szignifikáns pozitív korreláció (sőt szignifikáns negatív korrelációt lehetne kimutatni: a gyógyszer a férfiak és a nők esetében is hatékony, de a férfiak esetében kisebb arányban, ezért tűnhet úgy, mintha összességében hátráltatná a gyógyulást).

5. Kőzeteket csoportosítottunk aszerint, hogy tartalmaznak-e kvarcot, illetve plagioklász földpátot. Negyven vizsgált mintából huszonnyolcban volt plagioklász földpát, közülük tizennyolc tartalmazott kvarcot is. A plagioklászt nem tartalmazó minták között hét tartalmazott kvarcot.

Állíthatjuk-e $\alpha=0,05$ terjedelem (szignifikanciaszint) mellett, hogy a plagioklász földpát és a kvarc előfordulása között szignifikáns pozitív korreláció van?

Az alábbi táblázatot készíthetjük el:

	tartalmaz plagioklászt	nem tartalmaz plagioklászt	összesen
tartalmaz kvarcot	18	7	25
nem tartalmaz kvarcot	10	5	15
összesen	28	12	40

Minden osztályba esik legalább 4 megfigyelés, alkalmazható a pozitív korreláció tesztelésére vonatkozó módszer. A hipotézisek (a H_1 ellenhipotézis jelenti a pozitív korrelációt):

H_0: $\mathbb{P}$(tartalmaz kvarcot és plagioklászt is) $\leq \mathbb{P}$(tartalmaz kvarcot) $\cdot\, \mathbb{P}$(tartalmaz plagioklászt)

H_1: $\mathbb{P}$(tartalmaz kvarcot és plagioklászt is) $> \mathbb{P}$(tartalmaz kvarcot) $\cdot\, \mathbb{P}$(tartalmaz plagioklászt)

A pozitív korreláció ellenőrzéséhez az alábbi próbastatisztikát számítjuk ki:

$$u = \sqrt{n}\frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}} = \sqrt{40}\cdot\frac{18\cdot 5 - 10\cdot 7}{\sqrt{28\cdot 12\cdot 25\cdot 15}} = 0,36 < \Phi^{-1}(1-\alpha) = \Phi^{-1}(0,95) = 1,645.$$

Mivel $u < \Phi^{-1}(1-\alpha)$, elfogadható a nullhipotézis, nincs szignifikáns pozitív korreláció a kvarc és a plagioklász földpát jelenléte között.

6. Száztíz éven keresztül megfigyelték az éves csapadékösszeget és a maximális hőmérsékletet Budapesten (forrás: Országos Meteorológiai Szolgálat). Az alábbi táblázat tartalmazza, hogy hogyan alakult a kiemelkedően meleg (az évi középhőmérséklet legalább 12 fok), illetve kiemelkedően száraz (az éves csapadékösszeg legfeljebb 450 mm) évek száma.

csapadék/középhőmérséklet	$\geq 12°$C	$< 12°$C
< 450 mm	4	9
≥ 450 mm	8	89

Ezek alapján állíthatjuk-e $\alpha = 0,01$ terjedelem (szignifikanciaszint) mellett, hogy a kiemelkedően meleg időjárás és a kiemelkedően száraz időjárás előfordulása között szignifikáns pozitív összefüggés van?

Minden osztályba esik legalább 4 megfigyelés, elvégezhetjük a pozitív korrelációra vonatkozó próbát. A hipotézisek:

H_0: $\mathbb{P}(\geq 12$ fok és < 450 mm$) \leq \mathbb{P}(\geq 12$ fok$) \cdot \mathbb{P}(< 450$ mm$)$

H_1: $\mathbb{P}(\geq 12$ fok és < 450 mm$) > \mathbb{P}(\geq 12$ fok$) \cdot \mathbb{P}(< 450$ mm$)$

$r = s = 2$, mert mindkét szempont szerint két osztály van.

$N_{11} = 4;\ N_{12} = 9;\ N_{21} = 8;\ N_{22} = 89;\ n = 110$

$N_{1.} = 13; N_{2.} = 97; N_{.1} = 12; N_{.2} = 98$

$$u = \sqrt{n}\frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}} = \sqrt{110}\cdot\frac{4\cdot 89 - 9\cdot 8}{\sqrt{13\cdot 97\cdot 12\cdot 98}} = 2,45 > \Phi^{-1}(1-\alpha) = \Phi^{-1}(0,99) = 2,33.$$

ahol a kritikus érték a standard normális eloszlásfüggvényének inverze az $1-\alpha$ helyen. Tehát elutasítjuk a nullhipotézist, szignifikáns pozitív korreláció van a kiemelkedően meleg és a kiemelkedően száraz időjárás előfordulása között.