6. feladatsor

Statisztika

2010. december 6. és 8.

1. Egy $n = 10$ szervert tartalmazó kiszolgáló minden szervere minden pillanatban $0 < p < 1$ valószínűséggel foglalt, a foglaltságok szerverenként függetlenek. Tehát a foglaltak száma $Binom(n, p)$ eloszlásúnak tekinthető. A p-t szeretnénk meghatározni, ehhez 10 mérést végeztünk: 2,3,2,5,4,6,3,1,0,1.

   A minta alapján határozzuk meg p legvalószínűbb értékét, azaz adjuk meg a p maximum likelihood becslését általában $Binom(n, p)$ (fix n esetén), majd a konkrét példában.

   Megoldás: Általában $Binom(n, p)$ eloszlás és N elemű, $\vec{k} = (k_1, \ldots, k_N)$ minta megvalósulás esetén a likelihood függvény:

   $$L_p(\vec{k}) = \prod_{i=1}^{N} \binom{n}{k_i} p^{k_i} (1-p)^{n-k_i},$$

   p függvényében keressük a globális maximumhelyet. A példában $n = N = 10$ és a $k_1, \ldots, k_{10}$ értékek adottak. Tudjuk, hogy a maximum felvétetik valamely $0 \leq p \leq 1$-re, biztosan nem $p = 0$ vagy 1 esetén, és L_p deriválható, így a maximumhelyen $L'_p = 0$.

   Könnyebb a likelihood függvény logaritmusát deriválni, mert akkor függvények összegét deriváljuk, amely könnyebb feladat, mint szorzatot deriválni. Tehát a loglikelihood-függvény:

   $$l_p(\vec{k}) = \log L_p(\vec{k}) = \sum_{i=1}^{N} \log \binom{n}{k_i} + \sum_{i=1}^{N} k_i \log p + \sum_{i=1}^{N} (n - k_i) \log(1-p).$$

   Ennek deriváltja egyenlő 0-val (lok. szélsőérték helyek meghatározása)

   $$\partial_p l_p(\vec{k}) = \sum_{i=1}^{N} \frac{k_i}{p} - \sum_{i=1}^{N} \frac{n - k_i}{1-p} = 0.$$

   Átrendezéssel a maximum hely $\widehat{\mu}$:

   $$\widehat{\mu} = \frac{\sum_{i=1}^{N} k_i}{nN}.$$

   A konkrét példában $p = 0,27$. (Sokszor igaz, hogy ha csak egy helyen 0 a derivált, akkor az automatikusan maximumhely.)

2. Adjuk meg az (a) exponenciális, (b) Poisson, (c) geometriai eloszlásból vett minták maximum likelihood becslését.

   Megoldás:

   (a) Ha $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ függetlenek, akkor λ maximum likelihood becslése $\lambda = 1/\sum_{i=1}^{n} X_i/n$.

   (b) Ha $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$ függetlenek, akkor λ maximum likelihood becslése $\lambda = \sum_{i=1}^{n} X_i/n$.

   (C) Ha $X_1, \ldots, X_n \sim \text{Geom}(p)$ függetlenek, akkor p maximum likelihood becslése $p = n/\sum_{i=1}^{n} X_i$.

3. Tegyük fel, hogy egy kérdőívvel a megkérdezettek jövedelmi viszonyait akarják felderíteni. A korábbi tapasztalatok szerint a magas jövedelműek 0,2 valószínűséggel alacsony jövedelműnek vallják magukat. Az alacsony jövedelműek csupán 0,1 valószínűséggel állítják, hogy ők a magas jövedelműek. Adjunk maximum likelihood becslést a tényleges θ arányra az alapján, hogy a beérkezett kérdőívek közül x szólt magas, $n - x$ pedig alacsony jövedelemről.

   Megoldás: Ha a magas jövedelműek valódi aránya θ, akkor egy megkérdezett ember

   $$p = 0,1 \cdot (1-\theta) + 0,8 \cdot \theta = 0,1 + 0,7\theta$$

valószínűséggel vallja magát magas jövedelműnek, és $1-p$ valószínűséggel alacsony jövedelműnek. A likelihood függvény:

$$L_\theta(x,n-x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (0,1+0,7\theta)^x (0,9-0,7\theta)^{n-x}.$$

A log-likelihood függvény deriváltja 0-val egyenlő a következő helyen:

$$\widehat{\theta} = \frac{x-0,1n}{0,7n}.$$

4. A családok jövedelmét egy olyan skálán mérjük, ahol $X=1$ a létminimumnak felel meg. Feltételezzük, hogy a jövedelem eloszlása az $f(x) = \frac{\theta}{x^{\theta+1}}$ $(x \geq 1)$ sűrűségfüggvénnyel adható meg. (Ez az úgynevezett Pareto-eloszlás). Adjunk maximum likelihood becslést a θ-ra, ha 10 véletlenszerűen választott család jövedelme alapján: 1,53, 2,76, 19,65, 4,16, 7,31, 1,21, 254,2, 5,45, 1,12, 1,63.

   Megoldás: Folytonos esetben a likelihood-függvényben a sűrűségfüggvények szorzata szerepel, a ML-becslést ilyenkor is deriválással kaphatjuk meg. A ML becslés $\theta = n/(\sum_{i=1}^n \ln X_i)$, ahol $n=10$ és az $X_1,\ldots,X_{10}$ értékek a jövedelmek. A konkrét példában $\theta = 0,63$.

5. Egy alkatrész élettartama exponenciális eloszlású θ/t várható értékkel, ha t hőmérsékleten működtetjük. Tegyük fel, hogy az n megfigyelést a különböző $t_1,\ldots,t_n$ hőmérsékleten végeztük és $x_1,\ldots,x_n$ élettartamokat figyeltünk meg. Adjunk maximum likelihood becslést θ-ra.

   Megoldás: Ha egy exponenciális eloszlás várható értéke θ/t, akkor paramétere t/θ. Tehát sűrűségfüggvénye:

$$f_\theta(x) = \frac{t}{\theta} e^{-x\frac{t}{\theta}}, \text{ if } x \leq 0.$$

   Ha $(x_1,t_1),(x_2,t_2),\ldots,(x_n,t_n)$ a mért élettartam adat az adott hőmérsékleten, akkor a likelihood függvény:

$$L_\theta(\vec{x},\vec{t}) = \prod_{i=1}^n \frac{t}{\theta} e^{-x_i\frac{t_i}{\theta}}$$

   Ennek logaritmusa:

$$l_\theta(\vec{x},\vec{t}) = \sum_{i=1}^n \log(t_i) - \sum_{i=1}^n \log(\theta) - \sum_{i=1}^n \frac{-x_i t_i}{\theta}$$

   θ-szertinti deriváltja egyenlő 0-val:

$$-\frac{n}{\theta} + \sum_{i=1}^n \frac{-x_i t_i}{\theta^2} = 0.$$

   Innen átrendezéssel a ML becslés:

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^n x_i t_i}$$

6. Egy város energiafogyasztása normális eloszlású ismeretlen μ várható értékkel és a korábbi tapasztalatok alapján ismert σ szórással. n napon keresztül végeztünk méréseket $x_1,\ldots,x_n$ eredménnyel, majd az $(n+1)$-edik naptól m napon keresztül át csak a város egyik kerületéből érkeztek adatok, ahol a fogyasztás várható értéke az egész város fogyasztásának a fele: $y_1,\ldots,y_m$ a kapott adatsor. Tételezzük fel, hogy a szórás itt is σ. Adjunk maximum likelihood becslést μ-re.

   Megoldás: Legyen $X_1,\ldots,X_n$ $\mathcal{N}(\mu,\sigma)$ eloszlású minta. A mért megvalósulás: $\vec{x}=(x_1,\ldots,x_n)$.

   Legyen $Y_1,\ldots,Y_m$ $\mathcal{N}\left(\frac{\mu}{2},\sigma\right)$ eloszlású minta. A mért megvalósulás: $\vec{y}=(y_1,\ldots,y_m)$.

   A σ paraméter ismert. A likelihood függvény (ahogy általában) az $X_1,\ldots,X_n,Y_1,\ldots,Y_m$ együttes sűrűségfüggvénye:

$$L_\mu(\vec{x},\vec{y}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \prod_{j=1}^m \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_j-\mu/2)^2}{2\sigma^2}}.$$

   Ennek logaritmusa:

$$l_\mu(\vec{x},\vec{y}) = \sum_{i=1}^n \log\frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} + \sum_{j=1}^m \log\frac{1}{\sqrt{2\pi}\sigma} - \sum_{j=1}^m \frac{(y_j-\mu/2)^2}{2\sigma^2}.$$

   Ennek μ-szerinti deriváltja 0-val egyenlő:

$$\partial_\mu l_\mu(\vec{x},\vec{y}) = \sum_{i=1}^n \frac{2(x_i-\mu)}{2\sigma^2} + \sum_{j=1}^m \frac{2\frac{1}{2}(y_j-\mu/2)}{2\sigma^2} = 0.$$

Innen átrendezéssek kapjuk a ML becslést:

$$\widehat{\mu} = \frac{4\sum_{i=1}^{n} x_i + 2\sum_{j=1}^{m} y_j}{4n+m}$$

7. Egy cukorgyárban a cukrot 1000 gramm névértékű zacskókba csomagolják. A gyártási technológiából eredően az egy zacskóba kerülő cukor szórása 50 gramm, a várható értéke azonban ismeretlen, jelölje m gramm. Megvizsgálunk 25 zacskót, és azt tapasztaljuk, hogy a bennük lévő cukor mennyisége átlagosan 986 gramm. Elfogadjuk-e 95%-os szinten az $m = 1000$ hipotézist az $m \neq 1000$ hipotézis ellenében? Mi lenne a helyzet, ha a szórás csak 20 gramm lenne?

   Megoldás: A várható értékre tesztelünk, és a szórás ismert, tehát u-próbát alkalmazunk. Mivel a $H_0 : m = 1000$ nullhipotézist a $H_1 : m \neq 1000$ ellenében teszteljük, ezért a próba kétoldali. A minta mérete $n = 25$, a nullhipotézisre $\mu_0 = 1000$. A próbastatisztika a következő:

$$u(X) = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{986 - 1000}{50} * \sqrt{25} = -1,4,$$

   a normális eloszlás megfelelő kvantilis pedig

$$u_{0,025} = \Phi^{-1}(0.975) = 1,96.$$

   Mivel $-1,96 < -1,4 < 1,96$, így a nullhipotézist elfogadjuk.

   Ha azonban $\sigma = 20$, akkor $u(X) = 3,5 > 1,96$, így ebben az esetben az $m \neq 1000$ hipotézist fogadjuk el.

8. Egy oldat sótartalmát mérjük laborban. 5 mérést végzünk, melyek során a sótartalomra rendre a következő értékek adódtak (gramm/liter): 7.7, 8.1, 7.7, 7.5, 7.0. Az oldatról előzetesen azt állították, sótartalma 7 g/l. Elfogadjuk-e ezt 95%-os szinten azon hipotézis ellenében, hogy a sótartalom nem 7 g/l?

   Megoldás: A várható értékre tesztelünk, és a szórás ismeretlen, tehát t-próbát alkalmazunk. Szükség van az átlagra és a szórásnégyzetre: $\overline{x} = 7,6$, $\sigma_5^* = 0,36$. A statisztika $t(X) = \frac{\overline{x}-\mu_0}{s_n^*}\sqrt{n} = 3,75$, ezt hasonlítjuk össze a t-eloszlás kétoldali 95%-os kvantilisével, amelynek értéke $t_{0,05} = 2,78$. $3,75 > 2,78$ miatt 95%-os szinten elutasítjuk a $\mu_0 = 7$ nullhipotézist.

9. Megvizsgálták, hogy 10 ember mekkora távot tudott 5 perc alatt lefutni. Ezután mindenki 3 napot diétázott, és így is megmérték a futásteljesítményt. Azt szeretnénk kideríteni, hogy a diéta befolyásolta-e a futásteljesítményt. Bizonyítható-e 95%-os szinten, hogy a diéta javított a teljesítményen?

Diéta előtt	1520	1830	1620	1740	1970	2130	1910	2000	1980	1900
Diéta után	1630	1810	1700	1800	1930	2100	1960	2160	2040	1970

   Megoldás: Két adatsor van, de ugyanazokra az alanyokra vonatkozik, így egymintás t-próbát alkalmazunk az eltérésekre. Az eltérések:

110	-20	80	60	-40	-30	50	160	60	70

   Itt $\overline{x} = 50$ és $s_n^* = 60,5$. A $H_0 : \mu_0 = 0$ hipotézist teszteljük a $H_1 : \mu_0 > 0$ ellenében, így a próba egyoldali. A statisztika $t(X) = \frac{\overline{x}-\mu_0}{s_n^*}\sqrt{n} = \frac{50-0}{60,5}\sqrt{10} = 2,61$

   Ezt hasonlítjuk össze az $n - 1 = 9$ szabadsági fokú t-eloszlás egyoldali $0,05$-kvantilisével: $t_{0,05}(9) = 1,83$. Mivel $2,61 > 1,83$, ezért a nullhipotézist elutasítjuk, így 95%-os szinten állítható, hogy a diéta javít a futásteljesítményen.

10. Egy gyárban azt vizsgálják, milyen módon lehetne növelni a munkások termelékenységét. Kétféle módot tesztelnek: (A) fizetésemelés, (B) munkakörülmények javítása. Két külön csoporton tesztelnek. Az alábbi két táblázat tartalmazza a termelékenység változását.

(A)

munkás	1	2	3	4	5	6
változás	1.1	0.2	-0.1	2.2	1.3	1.3

(B)

munkás	1	2	3	4	5	6	7	8	9	10
változás	1.3	2.5	1.2	0.8	0.3	1.9	3.2	2.4	2.2	3.2

   (a) Fogadjuk el vagy utasítsuk el 95%-os szinten azt a nullhipotézist, hogy a fizetésemelés nem változtatja a termelékenységet.

(b) Fogadjuk el vagy utasítsuk el 95%-os szinten azt a nullhipotézist, hogy a munkakörülmények javítása nem változtatja a termelékenységet.

(c) Fogadjuk el vagy utasítsuk el 95%-os szinten azt a nullhipotézist, hogy a munkakörülmények javítása nem növeli jobban a termelékenységet, mint a fizetésemelés.

Megoldás:

(a) Egymintás, kétoldali t-próbát alkalmazunk. $\overline{x} = 1$, $s_x^* = 0,72$, $t(X) = 4,41 > t_{0.05}(5) = 2,57$, így a fizetésemelés szignifikánsan megváltoztatja a termelékenységet.

(b) $\overline{y} = 1,9$, $s_y^* = 0,93$, $t(Y) = 6,41 > t_{0.05}(9) = 2,262$, így a munkakörülmények javítása szignifikánsan megváltoztatja a termelékenységet.

(c) Két különböző csoportról van szó, ezért kétmintás, egyoldali t-próbát alkalmazunk. Az előzőleg kiszámolt értékek alapján a statisztika ($n_1 = 6, n_2 = 10$):

$$t(X,Y) = \frac{\overline{x} - \overline{y}}{\sqrt{(n_1 - 1)s_x^{*2} + (n_2 - 1)s_y^{*2}}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} = 0,54,$$

ami kisebb, mint az $n_1 + n_2 - 2$ szabadsági fokú egyoldali kvantilis, melynek értéke $t_{0,05}(14) = 1,78$, így a munkakörülmények javítása nem növeli jobban a termelékenységet, mint a fizetésemelés.

11. Kétféle tápot tesztelnek, az egyiket 6 csirkén, a másikat pedig 8 (az előzőektől különböző) csirkén. A tesztelésnél azt vizsgálják, mekkora a testsúlynövekedés a tápszer nélküli állapothoz képest. Az eredmény:

A típusú táp

csirke	1	2	3	4	5	6
növekedés	2.1	1.2	1.4	2.2	0.4	1.7

B típusú táp

csirke	1	2	3	4	5	6	7	8
növekedés	1.7	2.2	1.1	1.8	2.5	0.9	1.6	1.7

Döntsük el kétmintás t-próba segítségével, hogy a két táp hatása egyformának tekinthető-e 95%-os szinten.

Megoldás: $\overline{x} = 1.5$, $s_x^* = 0.61$, $\overline{y} = 1.69$, $s_y^* = 1.38$.

$$t(X,Y) = \frac{|\overline{x} - \overline{y}|}{\sqrt{(n_1 - 1)s_x^{*2} + (n_2 - 1)s_y^{*2}}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} = 0.34 < t_{0,05}(6 + 8 - 2) = 2.18,$$

így a két táp hatása egyformának tekinthető.

12. Amikor az embereket megkérdezik, hogy mekkora a tömegük, gyakran mondanak a valóságosnál kisebb értékeket. Szeretnénk eldönteni az alábbi adathalmazról, hogy igazi mérésből származik, vagy az emberek megkérdezéséből nyerték. Azt a tényt fogjuk használni, hogy mérés esetén az utolsó számjegyek eloszlásának egyenletesnek kell lennie a $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ halmazon. Döntsünk $0,95$ szinten arról a hipotézisről, hogy mérésből származnak az adatok.

utolsó számjegy	0	1	2	3	4	5	6	7	8	9
mérések száma	35	4	4	3	4	24	2	4	8	2

1.) Határozzuk meg H_0 és H_1 hipotéziseket.

2.) Határozzuk meg a próbastatisztika értékét. Mekkora a szabadságfoka a próbastatisztikának?

3.) Határozzuk meg a 0,95 szignifikanciaszinthez tartozó kritikus értéket a χ^2 eloszlás táblázatából.

4.) Hasonlítsuk össze a kritikus értéket és a próbastatisztika értékét, majd döntsünk arról, hogy elvetjük-e H_0-t, vagy nem.

Megoldás: Illeszkedésviszgálatot végzünk.

H_0: a számjegyek eloszlása egyenletes a $0, 1, \ldots, 9$ értékek között,

H_1: nem egyenletes.

$n = 90$ elemű a minta, $r = 10$ kategória van; a nullhipotézis fennállása esetén az egyes kategóriák valószínűsége $p_1 = p_2 = \ldots p_{10} = 1/10$.

Illeszkedésvizsgálathoz a próbastatisztika értéke

$$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = 114,6$$

míg a kritikus érték az $r-1=9$ szabadsági fokú χ^2-eloszlás $0,05$-kvantilise, ami $16,7$. $114,6 > 16,7$ miatt elutasítjuk H_0-t. Tehát 95%-os szinten beláttuk, hogy emberek megkérdezéséből jöttek az adatok.

13. Egy tóban háromféle hal él: amur, makréla és ponty. Ottó bácsi, az öreg horgász azt súgja nekünk, hogy a tóban kétszer annyi a ponty, mint a makréla vagy az amur. Kifogtunk 60 halat; döntsük el ez alapján 95%-os szinten, hallgathatunk-e Ottó bácsira.

amúr	makréla	ponty
11	14	35

Megoldás: Illeszkedésvizsgálatot végzünk. Ha Ottó bácsinak igaza van, akkor az egyes halak aránya a tóban: $p_1 = 1/4$ amur, $p_2 = 1/4$ makréla és $p_3 = 0,5$ ponty. A következő statisztikát számítjuk ki:

$$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i},$$

ahol $r=3$ a fajták (kategóriák) száma, $\nu_1 = 11$, $\nu_2 = 14$, $\nu_3 = 35$ az egyes fajtákból kifogott halak száma, $n = 60$ pedig az összes kifogott hal száma. A konkrét számokat behelyettesítve a statisztika értéke $1,97$, ezt kell tesztelnünk az $r-1=2$ szabadsági fokú χ^2 eloszlás 95%-os kvantilise ellen, amelynek értéke $5,99$. Mivel $1,97 < 5,99$, így H_0-at elfogadjuk, azaz hallgathatunk Ottó bácsira.

14. Azt szeretnénk megtudni, hogy a bukósisak színe és a baleseti sérülések típusa között van-e összefüggés. Az utóbbi néhány év adatai alapján a következő táblázatot kaphatjuk:

	fekete	fehér	sárga/narancs
Kontroll (nem sérült)	491	377	31
Balesetes (sérült vagy meghalt)	213	112	8

$\varepsilon = 0,05$ elsőfajú hibavalószínűséggel döntsünk arról a hipotézisről, hogy a csoport (kontroll vagy balesetes) független a bukósisak színétől.

Megoldás: Homogenitásvizsgálatot végzünk, azaz arra vagyunk kíváncsiak, hogy a sérült illetve a nem sérült sisakok esetében ugyanaz-e a színek eloszlása. A megfelelő statisztika

$$nm \sum_{i=1}^{r} \frac{(\nu_i/n - \mu_i/m)^2}{\nu_i + \mu_i},$$

ahol $r = 3$ a kategóriák száma, $\nu_1 = 491, \nu_2 = 377, \nu_3 = 31$ illetve $\mu_1 = 213, \mu_2 = 112, \mu_3 = 8$ az egyes kategóriákba eső megfigyelések száma és $n = 491 + 377 + 31 = 899$ és $m = 213 + 112 + 8 = 333$ az összes megfigyelések száma. A konkrét példában a statisztika értéke $8,77$, ezt teszteljük az $r-1=2$ szabadsági fokú χ^2 eloszlás 95%-os kvantilise ellen, amelynek értéke $5,99$. Mivel $8,77 > 5,99$, ezért elutasítjuk a nullhipotézist, és a különböző színű sisakok nem ugyanannyira védenek.

15. Egy csavar lehet hibás méret illetve szilárdság alapján is. Megvizsgáltunk 460 darab csavart.

	jó méret	rossz méret
jó szilárdság	416	16
rossz szilárdság	23	5

Döntsük el 95%-os szignifikanciaszinten, hogy az, hogy egy csavarnak megfelelő-e a szilárdsága illetve a mérete, független-e.

Megoldás: Függetlenségvizsgálatot végzünk. A táblázat megfelelő eleme legyen ν_{ij}. Ki kell számolnunk a peremeloszlásokat:

416	16	432
23	5	28
439	21	$n = 460$

A peremeloszlások: $p_1 = 432/460$, $p_2 = 28/460$ illetve $q_1 = 439/460$, $q_2 = 21/460$. A statisztika:

$$\sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(\nu_{ij} - np_i q_j)^2}{np_i q_j} = 12,09,$$

amit össze kell hasonlítanunk az $(r-1)(s-1) = 1$ (r és s az egyik változó, illetve a másik változó kategóriáinak száma, $r = 2$, $s = 2$) szabadsági fokú χ^2 eloszlás 95%-os kvantilisével, amelynek értéke $3,841$. Mivel $12,09 > 3,841$, ezért elutasítjuk a nullhipotézist, azaz a szilárdság illetve a méret nem függetlenek.