Statisztika **1. zárthelyi dolgozat** 2010. március 16.

1. A következő minta ismeretlen várható értékű, 1 szórású normális eloszlásból származik:

   2,3 5,1 3,8 2,7 3,1 3,9 4,2 2,5 3,4

   a) Számítsuk ki a mintaközepet, a mediánt és a korrigált tapasztalati szórást! (10 pont)
   b) Adjunk 90 %-os megbízhatóságú konfidenciaintervallumot a várható értékre! (10 pont)

2. $X_1, X_2, \ldots, X_n$ független minta $b > 0$ paraméterű Poisson-eloszlásból. Mi a b paraméter maximum likelihood becslése? (12 pont)

3. Ismeretlen r rendű és p paraméterű binomiális eloszlásból származó független minta tapasztalati közepe 10, tapasztalati szórása 3. Adjunk a paraméterekre becslést momentum-módszerrel. (8 pont)

4. $X_1, X_2, \ldots, X_n$ független minta az $(1, a)$ intervallumon egyenletes eloszlásból, ahol $a > 1$ ismeretlen paraméter. Milyen c valós számra lesz $T(X_1, \ldots, X_n) = 12s_n^{*2} + 4\overline{X} + c$ torzítatlan becslése az a paraméternek / az a paraméter négyzetének? (10 pont)

   Segítség: az l hosszúságú intervallumon egyenletes eloszlású valószínűségi változó szórása: $l/\sqrt{12}$.

A megoldásokat indokolni kell, a teljes pontszámhoz jó végeredmény és helyes indoklás szükséges. Összesen 50 pontot lehet elérni, az egyes feladatokért kapható pontszámok a feladat szövege mellett szerepelnek. A várható ponthatárok: 40, 61, 72, 83.

Az elégséges határa 20 pont, aki ezt nem éri el vagy a dolgozatot nem írja meg, szorgalmi időszak utolsó hetében vagy a vizsgaidőszak első hetében pótolhatja a dolgozatot.

Az eredmények az ETR infosheet rovatában lesznek elérhetők, a megoldásokat pedig a http://www.cs.elte.hu/~agnes/gyak címen lehet majd megtalálni.

1. A következő minta ismeretlen várható értékű, 1 szórású normális eloszlásból származik:

2,3 5,1 3,8 2,7 3,1 3,9 4,2 2,5 3,4

a) Számítsuk ki a mintaközepet, a mediánt és a korrigált tapasztalati szórást! (10 pont)
b) Adjunk 90 %-os megbízhatóságú konfidenciaintervallumot a várható értékre! (10 pont)

a) Jelölje a mintát $X_1, \ldots, X_9$. A minta elemszáma $n = 9$. A definíciók alapján számolunk.

A mintaközép: $\overline{X} = \frac{1}{9}(X_1 + \ldots + X_9) = \frac{1}{9} \cdot 31 \approx 3,44$.

A $n = 9$ elem esetén a medián a nagyság szerinti sorban a középső elem, $X_5^* = 3,4$.

A korrigált tapasztalati szórásnégyzet és szórás:

$$s_n^{*2} = \frac{9}{8}\left(\frac{1}{9}\sum_{i=1}^{9} X_i^2 - \overline{X}^2\right) = \frac{9}{8}\left(\frac{113,3}{9} - 3,44^2\right) \approx 0,81 \Rightarrow s_n^* \approx 0,9.$$

b) Normális eloszlás, ismert σ szórás esetén az előadáson tanultak szerint $1-\alpha$ megbízhatóságú konfidenciaintervallum a várható értékre:

$$\left(\overline{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

ahol $\Phi\left(u_{1-\alpha/2}\right) = 1 - \alpha/2$, és Φ a standard normális eloszlásfüggvény. Most 0,9 megbízhatóságú konfidenciaintervallumot szeretnénk, azaz $\alpha = 0,1$, továbbá $1 - \alpha/2 = 0,95$. A standard normális eloszlásfüggvény táblázata szerint $\Phi(1,64) \approx 0,9495$ és $\Phi(1,65) \approx 0,9505$, így $u_{1-\alpha/2} \approx 1,645$. $\overline{X} \approx 3,44$ az *a)* feladat szerint, a feladat szövege szerint $\sigma = 1$, és $n = 9$ a mintaelemek száma. Tehát a következő intervallum megfelelő:

$$\left(3,44 - 1,645 \cdot \frac{1}{3}; 3,44 + 1,645 \cdot \frac{1}{3}\right) = (2,89; 3,99).$$

2. $X_1, X_2, \ldots, X_n$ független minta $b > 0$ paraméterű Poisson-eloszlásból. Mi a b paraméter maximum likelihood becslése? (12 pont)

A Poisson-eloszlás diszkrét eloszlás, méghozzá $P(\xi = k) = \frac{b^k}{k!}e^{-b}$ $(k = 0, 1 \ldots)$, ha ξ b paraméterű Poisson-eloszlású valószínűségi változó. Ha ξ független a mintától, akkor segítségével a likelihood-függvény így írható:

$$f_b(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_b(\xi = X_i) = \prod_{i=1}^{n}\left[\frac{b^{X_i}}{X_i!}e^{-b}\right] = \left(\prod_{i=1}^{n}\frac{1}{X_i!}\right) b^{\sum_{i=1}^{n} X_i} e^{-nb}.$$

Észrevehetjük, hogy a mintaelemek összege elégséges statisztika.

Azt a b-t keressük, melyre a likelihood-függvény maximális. A likelihood-függvény most mindenképpen pozitív, és a logaritmusfüggvény szigorúan monoton növő, így elegendő a log-likelihood függvény maximumhelyét megkeresni:

$$l(b) = \log f_b(X_1, \ldots, X_n) = \log\left(\prod_{i=1}^{n}\frac{1}{X_i!}\right) + \sum_{i=1}^{n} X_i \log b - nb.$$

Ennek b szerinti deriváltja: $\partial l(b) = \sum_{i=1}^{n} X_i \cdot \frac{1}{b} - n$.

Könnyen látható, hogy ez pontosan akkor nulla, ha $b = \sum_{i=1}^{n} X_i \cdot \frac{1}{n} = \overline{X}$. Sőt $b < \overline{X}$ esetén a derivált pozitív, míg $b > \overline{X}$ esetén negatív, így a log-likelihood függvénynek valóban maximuma van a $b = \overline{X}$ helyen.

Tehát a $\hat{b} = \overline{X}$ becslés a paraméter maximum likelihood becslése.

3. Ismeretlen r rendű és p paraméterű binomiális eloszlásból származó független $X_1, \dots, X_n$ minta tapasztalati közepe 10, tapasztalati szórása 3 ($r \in \mathbb{N}, 0 < p < 1$). Adjunk a paraméterekre becslést momentum-módszerrel. (8 pont)

   Az első két momentumra írjuk fel, hogy a mintából kapott tapasztalati momentum megegyezik annak az eloszlásnak a megfelelő momentumával, melyből a minta származik. Felhasználjuk, hogy az r rendű és p paraméterű binomiális eloszlás várható értéke rp, szórásnégyzete pedig $rp(1-p)$, majd megoldjuk az egyenletrendszert.

$$\overline{X} = E_{r,p}(X_1) = rp, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = E_{r,p}\left(X_1^2\right) = D_{r,p}^2(X_1) + E_{r,p}(X_1) = rp(1-p) + r^2p^2;$$

$$\overline{X} = rp, \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = rp(1-p) \Rightarrow s_n^2 = rp(1-p);$$

$$\overline{X} = rp, \qquad 1 - p = \frac{s_n^2}{\overline{X}}.$$

   Ezekből kapjuk a megoldásokat:

$$\hat{p} = 1 - \frac{s_n^2}{\overline{X}} = 1 - \frac{3^2}{10} = 0,1; \qquad \hat{r} = \frac{\overline{X}}{1 - \frac{s_n^2}{\overline{X}}} = \frac{10}{0,1} = 100.$$

4. $X_1, X_2, \dots, X_n$ független minta az $(1, a)$ intervallumon egyenletes eloszlásból, ahol $a > 1$ ismeretlen paraméter. Milyen c valós számra lesz $T(X_1, \dots, X_n) = 12s_n^{*2} + 4\overline{X} + c$ torzítatlan becslése az a^2-nek? (10 pont)

   Tudjuk, hogy a korrigált tapasztalati szórásnégyzet torzítatlan becslése a szórásnégyzetnek, vagyis

$$E_a\left(s_n^{*2}\right) = D_a^2(X_1) = \frac{(a-1)^2}{12}.$$

   A mintaátlag pedig torzítatlan becslése a várható értéknek, vagyis

$$E_a\left(\overline{X}\right) = E_a(X_1) = \frac{a+1}{2}.$$

   Így a várható érték linearitása miatt (felhasználva még, hogy konstans várható értéke saját maga):

$$E_a(T(X_1, \dots, X_n)) = E_a\left(12s_n^{*2} + 4\overline{X} + c\right) = 12E_a\left(s_n^{*2}\right) + 4E_a\left(\overline{X}\right) + c =$$
$$= (a-1)^2 + 2(a+1) + c = a^2 + 3 + c.$$

   A T akkor torzítatlan becslése a^2-nek, ha minden lehetséges a paraméterre

$$E_a(T(X_1, \dots, X_n)) = a^2,$$

   a fentiek szerint látható, hogy ez akkor és csak akkor teljesül, ha $c = -3$. Ugyanígy látható, hogy egyetlen c valós számra sem teljesül, hogy T torzítatlan becslése az a paraméternek.

1. Egy étteremben minőségellenőrzéskor akkor tekintik a szolgáltatást megfelelőnek, ha a $H_0 : p = 0,8$ nullhipotézis elfogadható a $H_1 : p < 0,8$ ellenhipotézissel szemben, ahol p annak valószínűsége, hogy egy vendég elégedett volt (feltételezzük, hogy az egyes vendégek véleménye egymástól független, azonos eloszlású). Próbaként 20 embert kérdeznek meg, és akkor fogadják el a nullhipotézist, ha közülük legalább 19 elégedett volt. Mennyi az elsőfajú hiba valószínűsége? (8 pont)

2. Egy mérőállomáson az évi csapadékmennyiség sokéves átlaga 751 mm. Az utóbbi nyolc évben a következő értékek adódtak (szintén mm-ben):

   765 728 771 752 748 739 759 737

   Nullhipotézisünk az, hogy az utóbbi években a csapadékmennyiség várható értéke megegyezik a sokéves átlaggal, ellenhipotézisünk az, hogy eltér tőle. Elfogadható-e a nullhipotézis 0,05 elsőfajú hibavalószínűség mellett, ha a csapadékmennyiséget normális eloszlású valószínűségi változónak tekintjük? (10 pont)

3. Megmértük öt fizikus és öt matematikus testmagasságát, centiméterben a következő eredményeket kaptuk:

fizikus	176	164	186	178	183
matematikus	173	197	196	181	185

   Tekintsük a testmagasságot mindkét esetben 8 szórású, normális eloszlású valószínűségi változónak. $0,04$ elsőfajú hibavalószínűség mellett bizonyítható-e, hogy a matematikusok magasabbak a fizikusoknál? (10 pont)

4. Egy dobókockával százszor dobtunk, és feljegyeztük, hogy melyik szám hányszor jött ki:

érték	1	2	3	4	5	6
gyakoriság	19	16	23	15	13	14

   $0,01$ terjedelem mellett elfogadható-e, hogy a dobókocka szabályos, azaz minden szám egyforma valószínűséggel jön ki? (10 pont)

5. Egy újszülött testtömegét mérték életének első öt napján:

életkor (nap)	1	2	3	4	5
testtömeg (kg)	2,75	2,9	3,01	3,14	3,2

   a) Számístuk ki a regressziós egyenes egyenletét. (5 pont)
   b) Számítsuk ki az egyenes megbízhatóságát, R^2-t. (5 pont)
   c) Adjunk előrejelzést az újszülött tömegére 7 napos korára. (2 pont)

A megoldásokat indokolni kell, a teljes pontszámhoz jó végeredmény és helyes indoklás szükséges. Összesen 50 pontot lehet elérni, a várható ponthatárok: 40, 61, 72, 83.

Az elégséges határa 20 pont, aki ezt nem éri el vagy a dolgozatot nem írja meg, május 11-én 4 órától vagy a vizsgaidőszak első hetében (csütörtök vagy péntek, időpont később) pótolhatja a dolgozatot.

Az eredmények az ETR infosheet rovatában lesznek elérhetők, a megoldásokat pedig a http://www.cs.elte.hu/~agnes/gyak címen lehet majd megtalálni. Pótzh-ról és jegybeírásról minden tudnivaló a kurzusfórumon lesz olvasható.

1. Az elsőfajú hiba az, amikor a nullhipotézis igaz, de elvetjük. Vagyis a kérdés, hogy ha mindenki egymástól függetlenül $p = 0,8$ valószínűséggel elégedett, akkor mennyi annak valószínűsége, hogy húszból legalább legfeljebb 18-an elégedettek. Tudjuk, hogy ilyenkor az elégedett emberek száma, X binomiális eloszlású 20 renddel és $p = 0,8$ paraméterrel. Ezt felhasználva kapjuk, hogy az elsőfajú hiba valószínűsége:

$$P_0(X \le 18) = 1 - P_0(X \ge 19) = 1 - P_0(X = 20) - P_0(X = 19) =$$
$$= 1 - 0,8^{20} - 20 \cdot 0,8^{19} \cdot 0,2 \approx 0,9337.$$

   Megjegyzés: ez nagyon nagy érték, azt mutatja, hogy a próba elég rossz, ami abból is látszik, hogy a nullhiptézis mellett az elégedett emberek számának várható értéke 16, mégis csak legalább 19 elégedett ember esetén fogadjuk el a nullhipotézist.

2. Egyetlen normális eloszlásból származó mintánk van: $X_1, \ldots, X_n \sim N(m, \sigma)$ függetlenek, m valós, σ pozitív paraméter, mindkettő ismeretlen. A hipotézisvizsgálati feladat, milliméterben számolva:
   $H_0 : m = 751$;
   $H_1 : m \neq 751$.
   Erre tehát egymintás, kétoldali t-próbát végezhetünk. A próbastatisztika:

$$t = \frac{\overline{X} - m_0}{s_n^*}\sqrt{n} \approx \frac{749,875 - 751}{14,78}\sqrt{8} \approx -0,2153.$$

   A szabadsági fok $n - 1 = 7$, a terjedelem 0,05, így a kritikus érték a táblázat alapján 2,365. Kétoldali próbánál akkor vetjük a nullhipotézist, ha $|t| > c_{krit}$. Most $|t| \le c_{krit}$, így a nullhipotézist elfogadjuk. Az utóbbi években a csapadékmennyiség várható értéke nem tér el a sokévi átlagtól, vagy kevés az adat ennek cáfolatához.

3. Két normális eloszlású mintánk van, világos, hogy ezek egymástól is függetlenek: $X_1, \ldots, X_n \sim N(m_1, 8)$ a fizikusok, $Y_1, \ldots, Y_n \sim N(m_2, 8)$ a matematikusok testmagassága. A feladat centiméterben számolva:
   $H_0 : m_1 = m_2$;
   $H_1 : m_1 < m_2$.
   Kétmintás, egyoldali u-próbát végezhetünk, hiszen független, normális eloszlású mintákról van szó, és a szórások ismertek. A próbastatisztika:

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{177,4 - 186,4}{\sqrt{\frac{64}{5} + \frac{64}{5}}} \approx -1,778.$$

   A kritikus érték a standard normális eloszlásfüggvény táblázata alapján:

$$\Phi^{-1}(0,04) = -\Phi^{-1}(1 - 0,04) = -\Phi^{-1}(0,96) \approx -1,75.$$

   Akkor vetjük el a nullhipotézist, ha u kisebb a kritikus értéknél. Most ez teljesül, a nullhipotézist elvetjük, az adott terjedelem mellett az adatok statisztikai értelemben bizonyítják, hogy a matematikusok magasabbak a fizikusoknál.

   Megjegyzés: a próbastatisztika és kritikus érték közel van egymáshoz, hasonló helyzetben érdemes más terjedelmet választani.

4. Legyen A_i az az esemény, hogy a dobókockával i-t dobunk ($i = 1, \ldots, 6$). A nullhipotézis az, hogy a dobókocka szabályos, azaz $P(A_i) = p_i = 1/6$ minden i-re, az ellenhipotézis, hogy ettől eltérő az eloszlás. Ez illeszkedésvizsgálati feladat, χ^2-próbát végezhetünk, minden osztályba

esik legalább öt megfigyelés. Az osztályok száma: $r = 6$, a megfigyelések száma: $n = 100$. A gyakoriságokat ν_i-vel jelölve a próbastatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = \sum_{i=1}^{6} \frac{(\nu_i - 100/6)^2}{100/6} \approx 4,16.$$

A szabadsági fok: $f = r - 1 = 5$, a terjedelem $0,01$, a kritikus érték a táblázat alapján $15,1$. χ^2-próbánál akkor utasítjuk el a nullhipotézist, ha a próbastatisztika a kritikus értéknél nagyobb. Most tehát a nullhipotézist elfogadjuk, az adott terjedelem mellett elfogadható, hogy a dobókocka szabályos, vagy valóban szabályos, vagy kevés az adat ennek cáfolatához.

5. Legyen $X_i = i$, Y_i az újszülött tömege i napos korában, kilogrammban ($i = 1, 2, \ldots$). Az $Y_i = aX_i + b + \varepsilon_i$ lineáris modellt tekintjük, ahol ε_i-k egymástól független, nulla várható értékű, azonos szórású valószínűségi változók. Az ismert képleteket használjuk a továbbiakban. Megfigyeléseink száma: $n = 5$, mindkét mintaközép 3.
a) A lineáris együttható becslése:

$$\hat{a} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \approx 0,114.$$

A konstans tag becslése:

$$\hat{b} = \overline{Y} - \hat{a}\overline{X} = 3 - \hat{a} \cdot 3 \approx 2,658.$$

Tehát a regressziós egyenes egyenlete: $y = 0,114x + 2,658$.
b)

$$R^2 = \frac{\left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \cdot \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \approx 0,983.$$

Ez elég közel van az egyhez, azaz az adatok jól közelíthetők a lineáris modellel.
c) $\hat{Y}_7 = \hat{a}X_7 + \hat{b} \approx 0,114 \cdot 7 + 2,658 \approx 3,456$. Az előrejelzés szerint az újszülött 7 napos korára 3,46 kg tömegű lesz.