Valószínűségszámítás és statisztika
11. gyakorlat
2014. 12. 04.

Információ

ZH

Jövő héten (2014.12.11-én) ZH. Helyszín: Déli Tömb 0.803. (Szabó József terem).
Csak minden páratlanadik sorba üljetek, és egy sorba 5 ember üljön (természetesen itt is maradjon ki legalább 1-2 szék).
Legyetek szívesek névsor szerint helyet foglalni (a táblától nézve lentről felfelé, balról jobbra folytonosan):
1. sor: Á-B; 3. sor: Cs-F; 5. sor: G-K; 7. sor: L-Né; 9. sor: Ny-Sz; 11. sor: T-V.
Lehetőleg 14:05-re foglaljátok el a helyeiteket. Minél hamarabb kész vagytok, annál több időtök marad a ZH-ra.
Mindenkinél legyen fényképes igazolvány (diákigazolvány, személyi igazolvány, jogosítvány, útlevél).
Ezen a ZH-n csak egy A5-ös saját kézzel írott puska, csak a feladatsorokat tartalmazó egy darab nyomtatott A4-es lap, a statisztikai táblázat és nem programozható számológép használható. Jegyzet, könyv, tablet, telefon, laptop stb. nem. Az egymással és a külső személyekkel való kommunikáció semmiféle formában nem engedélyezett a ZH ideje alatt.
Mindenki hozzon magával legalább 5 üres lapot. Előttetek csak üres lap, az A5-ös puska, a feladatsorokat tartalmazó A4-es lap, a statisztikai táblázat, számológép, íróeszköz, fényképes igazolvány és étel-ital maradhat. A kabátokat és a táskákat lehetőség szerint hagyjátok a ruhatárban, illetve az üresen maradt sorok padjaira pakoljátok őket.
Minden lapon szerepeljen a nevetek és a neptunkódotok. Lehetőleg minden feladatot külön oldalra írjatok. Törekedjetek a világos, jól olvasható leírásra. Csak arra tudok pontot adni, amit el tudok olvasni. A puszta eredményközlés nem sokat ér, a teljes, hibátlan levezetés ér maximális pontot (természetesen részpontszám is kapható). A ZH-t hosszában hajtsátok ketté (a név legyen kifelé), és úgy adjátok majd be.

Konzultáció

Jövő héten kedden (2014.12.09-én) 17:15-től ZH előtti konzultáció lesz. Helyszín: Déli Tömb 00.112.

Elmélet

Hipotézis $\sim$ valami állítás, aminek igazságát vizsgálni szeretnénk Paramétertér: $\Theta = \Theta_0 \cup^* \Theta_1 \longrightarrow$ "valóság"

Mintatér: $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k \longrightarrow$ "látszat" - MINTÁBÓL
$\mathcal{X}_k$: kritikus tartomány - azon **X** megfigyelések halmaza, amikre *elutasítjuk* a nullhipotézist
$\mathcal{X}_e$: elfogadási tartomány - azon **X** megfigyelések halmaza, amikre *elfogadjuk* a nullhipotézist

Hipotézisvizsgálati feladat:
$H_0: \ \theta \in \Theta_0$ $\rightsquigarrow$ nullhipotézis
$\underline{H_1: \ \theta \in \Theta_1}$ $\rightsquigarrow$ ellenhipotézis

Tehát ha $\mathbf{X} \in \mathcal{X}_e$, akkor elfogadjuk H_0-t; ha $\mathbf{X} \in \mathcal{X}_k$, akkor pedig elutasítjuk H_0-t.
Amennyiben a Θ_0 halmaz egyelemű, akkor azt mondjuk, hogy H_0 egyszerű. H_1-re ugyanígy.

Az $\mathcal{X}$ mintatér felosztását általában egy statisztika (neve: próbastatisztika) segítségével végezzük el:
legyen T: $\mathcal{X} \to \mathbb{R}$, $\mathcal{X}_k = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) > c\}$ c neve: kritikus érték
$\mathcal{X}_e = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) \leq c\}$

döntés / "valóság"	H_0-t elfogadjuk ($\mathcal{X}_e$)	H_0-t elutasítjuk ($\mathcal{X}_k$)
H_0 teljesül (Θ_0)	helyes döntés	elsőfajú hiba
H_0 nem teljesül (Θ_1)	másodfajú hiba	helyes döntés

P(elsőfajú hiba)=$\alpha(\theta)$=$\mathrm{P}_\theta(\mathcal{X}_k)$, ahol $\theta \in \Theta_0$
P(másodfajú hiba)=$\beta(\theta)$=$\mathrm{P}_\theta(\mathcal{X}_e)$, ahol $\theta \in \Theta_1$
Erőfüggvény: ψ: $\Theta_1 \to \mathbb{R}$, $\psi(\theta) = P_\theta(\mathcal{X}_k)$
Terjedelem: α = sup $\{\alpha(\theta)$: $\theta \in \Theta_0\}$

Azt mondjuk, hogy az 1-es próba *erősebb* a 2-es próbánál, ha $\alpha_1 = \alpha_2$ és $\psi_1(\theta) \geq \psi_2(\theta)\ \forall \theta \in \Theta_1$.
Próbafüggvény: φ: $\mathcal{X} \to$[0,1] $\rightsquigarrow$ ennyi valószínűséggel vetem el a H_0-t a minta alapján
$\mathbf{x} \in \mathcal{X}_k \Rightarrow \varphi(\underline{x}) = 1$
$\mathbf{x} \in \mathcal{X}_e \Rightarrow \varphi(\underline{x}) = 0$
p-érték: az az α terjedelem, ami esetén a próbastatisztika értéke egyenlő a kritikus értékkel : T(**x**)= c_α.
A p-érték a legkisebb terjedelem, amire még elutasítjuk a H_0-t. Ha egy próbát számítógép segítségével végzünk el, rendszerint a p-érték révén tudunk dönteni: ha (p-érték)$< \alpha$, akkor elvetjük H_0-t.

Ha mind H_0, mind H_1 egyszerű, akkor adott α terjedelemhez lehet legerősebb próbát találni, ezt pedig úgy hívják, hogy *valószínűség-hányados próba*. A hipotéziseket folytonos esetre írom fel. Diszkrétre a sűrűségfüggvény helyett a konkrét eloszlást kell írni.
$H_0: \ f = f_0$
$\underline{H_1: \ f = f_1}$
A valószínűség-hányados próba kritikus tartománya: $\mathcal{X}_k = \left\{\mathbf{x} : \frac{f_1(\underline{x})}{f_0(\underline{x})} > c_\alpha\right\}$

Személyes adatok: Név: Bondici László E-mail: bondici@cs.elte.hu Honlap: www.cs.elte.hu/∼bondici

Tehát azokat az **x**-eket, amire az $\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ nagy, bepakoljuk a kritikus tartományba egészen addig, míg az adott α terjedelmet el nem érjük. Diszkrét esetben ehhez általában véletlenítésre van szükség, azaz bizonyos **x**-ek esetén nem 1 vagy 0, hanem egy, e két szám közé eső (jelöljük p_α-val) valószínűséggel vetjük el a nullhipotézist.

Néhány konkrét próba – az α végig a próba terjedelmét jelöli, ami előre adott

1.) Egymintás próbák

a.) Egymintás u-próba

$X_1, \ldots, X_n \sim$N(m, σ^2), ahol σ ismert, m paraméter

a.) $H_0 : m = m_0$, $H_1 : m \neq m_0$ b.) $H_0 : m = m_0$, $H_1 : m > m_0$ c.) $H_0 : m = m_0$, $H_1 : m < m_0$

A próbastatisztika: T(**X**)=$u = \sqrt{n}\frac{\overline{X}-m_0}{\sigma} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$

A kritikus tartományok:

a) $\mathcal{X}_k = \{\mathbf{x} : |u| > u_{\alpha/2}\}$

b) $\mathcal{X}_k = \{\mathbf{x} : u > u_\alpha\}$

c) $\mathcal{X}_k = \{\mathbf{x} : u < -u_\alpha\}$

b.) Egymintás t-próba

$X_1, \ldots, X_n \sim$N(m, σ^2), ahol σ, m paraméter

a.) $H_0 : m = m_0$, $H_1 : m \neq m_0$ b.) $H_0 : m = m_0$, $H_1 : m > m_0$ c.) $H_0 : m = m_0$, $H_1 : m < m_0$

A próbastatisztika: T(**X**)=$t = \sqrt{n}\frac{\overline{X}-m_0}{s_n^*} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_{n-1}$

A kritikus tartományok:

a) $\mathcal{X}_k = \{\mathbf{x} : |t| > t_{n-1,\alpha/2}\}$

b) $\mathcal{X}_k = \{\mathbf{x} : t > t_{n-1,\alpha}\}$

c) $\mathcal{X}_k = \{\mathbf{x} : t < -t_{n-1,\alpha}\}$

2.) Kétmintás próbák

$X_1, \ldots, X_n \sim$N(m_1, σ_1^2)

$Y_1, \ldots, Y_m \sim$N(m_2, σ_2^2)

Az elvégzendő próbák $H_0 : m_1 = m_2$ nullhipotézis esetén:

	a két minta független		a két minta nem független
σ_1 és σ_2 ismert	b.) kétmintás u-próba		egymintás u-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes F-próba		egymintás t-próba a különbségekre
	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	
	c.) kétmintás t-próba	d.) Welch-próba	

a.) F-próba

$m_1, m_2, \sigma_1, \sigma_2$ paraméterek

$H_0 : \sigma_1 = \sigma_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $F = \begin{cases} \frac{(s_1^*)^2}{(s_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1} & \text{ha } s_1^* > s_2^* \\ \frac{(s_2^*)^2}{(s_1^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{m-1,n-1} & \text{ha } s_2^* > s_1^* \end{cases}$

b.) kétmintás u-próba

m_1, m_2 paraméterek, σ_1, σ_2 ismert

$H_0 : m_1 = m_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $u = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \overset{H_0 \text{ esetén}}{\sim}$N(0,1)

c.) kétmintás t-próba

$m_1, m_2, \sigma_1 = \sigma_2$ paraméterek

$H_0 : m_1 = m_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $t = \sqrt{\frac{nm}{n+m}} \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(n-1)(s_1^*)^2+(m-1)(s_2^*)^2}{n+m-2}}} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_{n+m-2}$

d.) Welch-próba

$m_1, m_2, \sigma_1 \neq \sigma_2$ paraméterek

$H_0 : m_1 = m_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $t' = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(s_1^*)^2}{n}+\frac{(s_2^*)^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_f$, ahol

$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$

$c = \frac{\frac{(s_1^*)^2}{n}}{\frac{(s_1^*)^2}{n}+\frac{(s_2^*)^2}{m}}$, ha $s_1^* > s_2^*$

3.) χ^2-próbák

a.) Diszkrét illeszkedésvizsgálat

Feladat: adott egy $\mathbf{X} = (X_1, \ldots, X_n)$ n elemű minta, és azt akarjuk eldönteni, hogy a minta egy általunk "remélt" eloszlásból származik-e. *Diszkrét* illeszkedésvizsgálatnál feltesszük, hogy a mintaelemek r különböző értéket vehetnek fel: P($X_i = x_j$) $= p_j \quad j = 1, \ldots, r$. Jelöljük N_j-vel a gyakoriságokat, azaz azt, hogy az n elemű mintában hány darab x_j szerepel.

Osztályok	1	2	…	r	Összesen
Valószínűségek	p_1	p_2	…	p_r	1
Gyakoriságok	N_1	N_2	…	N_r	n

H_0 : a valószínűségek: **p=**$(p_1, \ldots, p_r)$

H_1: nem ezek a valószínűségek

A próbastatisztika: $T_n = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ eloszlásban, ha $n \to \infty$

A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valamilyen eloszlású, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbastatisztika: $T_n \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

b.) Függetlenségvizsgálat

Feladat: van egy minta, két szempont szerint csoportosítva. Azt kell eldönteni, hogy a két szempont független-e egymástól.

$\mathrm{p}_{i,j}$=P(egy megfigyelés az (i,j) osztályba kerül)

$\mathrm{N}_{i,j}$=ennyi megfigyelés kerül az (i,j) osztályba

A mintavétel eredménye:

		2. szempont					
		1	…	j	…	s	Összesen
1. szempont	1	N_{11}	…	N_{1j}	…	N_{1s}	$N_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
	i	N_{i1}	…	N_{ij}	…	N_{is}	$N_{i\bullet}$
	⋮	⋮		⋮		⋮	⋮
	r	N_{r1}	…	N_{rj}	…	N_{rs}	$N_{r\bullet}$
Összesen		$N_{\bullet 1}$	…	$N_{\bullet j}$	…	$N_{\bullet s}$	n

$N_{i\bullet} = \sum_{j=1}^{s} N_{i,j}$

$N_{\bullet j} = \sum_{i=1}^{r} N_{i,j}$

H_0 : a szempontok függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j}$ $\forall i, j$-re

H_1: nem azok

A próbastatisztika: $T_n = n \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{N_{i,j}^2}{N_{i\bullet} N_{\bullet j}} - 1 \right) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{(r-1)(s-1)}$ eloszlásban, ha $n \to \infty$

A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\underline{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$

Feladatok

1. (10-es feladatsor 6-os feladata) Legalább hány embert kell megkérdezni egy közvélemény-kutatásnál, ha egy adott párt támogatottságát legalább 95%-os valószínűséggel 0.01–nál kisebb eltéréssel szeretnénk megbecsülni? Számoljunk egyrészt a Csebisev-egyenlőtlenséggel, másrészt a normális eloszlással.

2. Valaki azt állítja, hogy a klíma változik, és ezt azzal véli bizonyítottnak, hogy az elmúlt 10 évben 2-szer is volt jégeső, pedig korábban az egyes évekre a jégeső valószínűsége a hivatalos adatok alapján csupán $p = 0.1$ volt. Írjuk fel a hipotéziseket, a próbát és állapítsuk meg az elsőfajú hiba valószínűségét, valamint az erőfüggvényt a $p = 0.2$ pontban!

3. Az alábbi minta 4 év október 18-án Budapesten mért napi középhőmérséklet adatait tartalmazza. Ellenőrizzük a $H_0 : m = 15$ hipotézist $\alpha = 0.05$ elsőfajú hibavalószínűség mellett a $H_1 : m < 15$ alternatívával szemben.

   a) A korábbi tapasztalatok alapján tekintsük az értékek szórását 2-nek. Adjuk meg a p-értéket is.

   b) Ne használjunk a szórásra vonatkozóan előzetes információt.

Középhőmérséklet (Celsius-fok)	14.8	12.2	16.8	11.1

4. A Dezinformatikai Kar III. évfolyamán 10-en írtak statisztika zárthelyit. 2 feladatsor volt, mindkettőben 30 pontot lehetett elérni. A pontszámokat tartalmazza az alábbi táblázat:

1. feladatsor	12	11	8	14	10
2. feladatsor	15	14	9	16	11

   Vajon az első feladatsor nehezebb volt? Mennyiben változik a helyzet, ha nem 10 diákról, hanem csak 5-ről van szó, és a 2. feladatsor a pótzh eredménye?

5. Értelmezzük az alábbi két számítógépes programot és eredményt:

   a)
   x=rnorm(100,0,1)
   t.test(x, alternative="t", mu=0.1)
   t = -0.9124, df = 99, p-value = 0.3638
   alternative hypothesis: true mean is not equal to 0.1
   95 percent confidence interval: -0.165 0.198
   sample estimates: mean of x 0.016

   b)
   x=rnorm(1000,0,1)
   t.test(x, alternative="t", mu=0.1)
   t = -2.7081, df = 999, p-value = 0.006882
   alternative hypothesis: true mean is not equal to 0.1
   95 percent confidence interval: -0.050 0.075
   sample estimates: mean of x 0.0126

6. A 10-es feladatsor 5-ös feladatában vizsgáljuk meg a kapott regressziós együtthatók szignifikanciáját.

7. A Dezinformatikai Kar III. évfolyamán 300-an tanulnak. Megszámolták, hogy a legutóbbi vizsgaidőszakban hányszor buktak az egyes hallgatók. Az eredményeket tartalmazza az alábbi táblázat:

Bukások száma	0	1	2	3	4
Hallgatók száma	80	113	77	27	3

   Elfogadhatjuk-e azt a hipotézist, hogy egy hallgató bukásszáma binomiális eloszlású $(4, 0.25)$ paraméterrel? És azt, hogy binomiális $(4, p)$ eloszlású?

8. Az alábbi kontingencia-táblázat mutatja, hogy 100 évben a csapadék mennyisége és az átlaghőmérséklet hogyan alakult (a cellákban az egyes esetek gyakoriságai találhatók):

Hőmérséklet \ Csapadék	Kevés	Átlagos	Sok
Hűvös	15	10	5
Átlagos	10	10	20
Meleg	5	20	5

   Tekinthető-e a csapadékmennyiség és a hőmérséklet függetlennek?