Matematikai statisztika

Informatika alapszak, "A" spec.

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Matematikai Intézet
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Honlap: zempleni.elte.hu

E-mail: andras.zempleni@ttk.elte.hu

Szoba: D 3-310

9. előadás

Nemparaméteres próbák

- Eddig végig feltettük a minta normalitását. Ez sokszor nem reális. Ha kiugró értékek vannak az adatsorban (vastag szélű eloszlásból származnak), akkor pl. a t-próba nem használható.
- Előjelpróba
  - Egymintás teszt a $P(X > m) = 1/2$ nullhipotézis vizsgálatára: Összeszámoljuk az m-nél nagyobb mintaelemeket, ez H_0 esetén Bin(n;1/2) eloszlású.
  - Párosított mintákra a $P(X > Y) = 1/2$ nullhipotézis vizsgálatára: Összeszámoljuk azokat a mintaelemeket, ahol $X_i > Y_i$, ez H_0 esetén Bin(n;1/2) eloszlású.
- Wilcoxon (Mann-Whitney) próba független mintákra a $P(X > Y) = 1/2$ nullhipotézis vizsgálatára: rangstatisztika (csak a sorbarendezett mintaelemek sorszámán múlik): $W = \sum_{i,j} I(X_i > Y_j)$. W aszimptotikusan normális eloszlású, H_0 esetén $EW = nm/2$, $D^2(W) = \frac{nm(n+m+1)}{12}$ ebből számolhatóak a kritikus értékek

A χ^2-próba

Legyen $A_1, \ldots, A_r$ teljes eseményrendszer.
Végezzünk n darab független megfigyelést, jelölje az i-edik esemény bekövetkezési gyakoriságát N_i $(i = 1, \ldots, r)$. A megfigyelések egyes eredményei segítségével definiálható az X_i valószínűségi változó, ami vegyen fel olyan értéket, amelyik számú esemény a teljes eseményrendszerből bekövetkezett. Ezáltal formálisan

$$N_i = \sum_{j=1}^{n} I(X_j = i) \text{ és } \sum_{i=1}^{r} N_i = n$$

H_0: $P(A_i) = p_i$, $i = 1, \ldots, r$ $\rightsquigarrow$ tfh. $p_i > 0 \ \forall i$, $p_1 + \ldots + p_r = 1$
H_1: a nullhipotézis tagadása

Próbastatisztika: $T_n(\mathbf{X}) := \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{r-1}$ eloszlásban

Kritikus tartomány: $\mathcal{X}_k = \left\{\mathbf{x} : T_n(\mathbf{X}) > \chi^2_{r-1;1-\alpha}\right\}$

A χ^2-próba II

Alkalmazásai:

- illeszkedésvizsgálat: egy minta adott eloszlást követ-e
- homogenitásvizsgálat: két minta eloszlása megegyezik-e
- függetlenségvizsgálat: két szempont, ismérv, tulajdonság független-e egymástól

Megjegyzések:

- a χ^2-próba **aszimptotikus** próba, ami azt jelenti, hogy "nagy" mintaelemszámra lehet végrehajtani. "Kicsi" minták esetén a kritikus érték nem használható, azt szimulálni kell a konkrét minta alapján.
- Mikor elég "nagy" már egy minta – hüvelykujjszabály: ha legalább 100 elemű. Egyébként H_0-tól függ, hogy legalább mekkora n-re van szükség, hogy kritikus értéknek a χ^2-eloszlás kvantiliseit lehessen használni.
- Végrehajtásának további feltétele, hogy minden osztályban "elegendő" mennyiségű gyakoriság legyen (szokásos feltétel: $N_i \geq 4$).
- A próbastatisztikában lévő összeg tagjai $\frac{(O-E)^2}{E}$ alakúak, ahol E: elméleti gyakoriságok, O: tapasztalati gyakoriságok

Illeszkedésvizsgálat

H_0: a minta egy adott eloszlásból származik
H_1: a minta nem ilyen eloszlású

Végrehajtása:

- grafikus módszerek ("szemmel" jónak tűnik-e az illeszkedés):
  - Q-Q plot
  - P-P plot
  - hisztogram/magfüggvényes sűrűségfüggvény-becslés, valamint az illesztett sűrűségfüggvény egy ábrában
- statisztikai próbák:
  - diszkrét eloszlás esetén χ^2-próba
  - folytonos eloszlás esetén több statisztikai próba közül lehet választani
    - diszkretizálás (mesterséges osztályok létrehozása) révén χ^2-próba
    - Kolmogorov-Szmirnov próba
    - Cramér-von Mises próba
    - Anderson-Darling próba
    - Shapiro-Wilk próba: kizárólag normalitásvizsgálatra, amire ez a legjobb

Illeszkedésvizsgálat χ^2-próbával

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: $\mathbf{p}=(p_1,\ldots,p_r)$

H_1: nem ezek a valószínűségek

Próbastatisztika: $T_n(\mathbf{X}) = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ elo.-ban, ha $n \to \infty$

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: $T_n(\mathbf{X}) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

A χ^2-próba végrehajtásának feltételei (hüvelykujjszabály): $N_i \geq 4$ és $np_i \geq 4$ minden i-re. Ha ezek nem teljesülnek, akkor osztályokat kell összevonni.

Illeszkedésvizsgálat Kolmogorov-Szmirnov próbával

$H_0 : F_{X_1}(x) = F(x) \ \ \forall x \in \mathbb{R}$ ahol F egy adott eloszlás elofv.-e
H_1: a nullhipotézis tagadása

Próbastatisztika: $D_n(\mathbf{X}) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$

A próbastatisztika $\sqrt{n}$-szeresének eloszlása H_0 esetén az ún. Kolmogorov-eloszláshoz tart ($n \to \infty$). Jelöljük K_α-val a Kolmogorov-eloszlás α-kvantilisét.

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : \sqrt{n} D_n(\mathbf{x}) > K_{1-\alpha}\}$

Megjegyzések:

- D_n kiszámításához elég csak a mintapontokban tekinteni az eltérést.
- Nem lehet használni a határeloszlást, ha paramétereket kell becsülnünk! Ilyen esetben a kritikus értéket szimulációval kaphatjuk meg.
- A Kolmogorov-eloszlás eloszlásfüggvénye: $1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}$

Feladatok

E10.) Egy gyártó megfigyelte, hogy 100, általa előállított SSD merevlemezen 5 év használat után hány hibás szektort talál az ezek felkutatására készített szoftver:

Hibás szektorok száma	0	1	2	3	4	5	7	Összesen
Gyakoriságok	45	35	12	5	1	1	1	100

Vizsgáljuk meg, hogy a szektorhibák száma Poisson-eloszlást követ-e!

E11.) Nézzük meg P-P plot-tal és Q-Q plot-tal, majd diszkretizálás után χ^2-próbával, valamint Kolmogorov-Szmirnov próbával, hogy a következő minta:

4,3 2,0 5,6 8,1 3,2 0,6 5,4 8,9 7,5 9,3
9,6 6,7 4,4 2,9 1,0 6,5 4,0 6,6 4,2 1,9

származhat-e az alábbi eloszlásokból:

a.) $E(0;10)$;

b.) $N\left(5;\left(\frac{5}{\sqrt{3}}\right)^2\right)$.

Homogenitásvizsgálat

Adott két független minta, mindkettő egy közös szempont szerint r osztály egyikébe sorolva.

	Osztályok	1	2	...	r	Összesen
1. minta	Valószínűségek	p_1	p_2	...	p_r	1
	Gyakoriságok	N_1	N_2	...	N_r	n
2. minta	Valószínűségek	q_1	q_2	...	q_r	1
	Gyakoriságok	M_1	M_2	...	M_r	m

H_0: a két minta azonos eloszlású, azaz $(p_1, \ldots, p_r) = (q_1, \ldots, q_r)$
H_1: a nullhipotézis tagadása

Próbastatisztika: $T_{n,m}(\mathbf{X}, \mathbf{Y}) = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{r-1}$ eloszlásban

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : T_{n,m}(\mathbf{X}, \mathbf{Y}) > \chi^2_{r-1,1-\alpha}\}$

Függetlenségvizsgálat

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

$p_{i,j} = P$(egy megfigyelés az (i,j) osztályba kerül)
$N_{i,j} =$ennyi megfigyelés kerül az (i,j) osztályba

		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

ahol $N_{i\bullet} = \sum_{j=1}^{s} N_{ij}$ és $N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$

Függetlenségvizsgálat II

Itt formálisan a mintánk két dimenziós: a megfigyelések az $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ párok, ahol az X-ek r, az Y-ok pedig s különböző értéket vehetnek fel nemnulla valószínűséggel: $p_{i,j} = P(X_1 = x_i, Y_1 = y_j)$, ahol $i = 1, \ldots, r$ és $j = 1, \ldots, s$.

Továbbá $N_{i,j} = \sum_{k=1}^{r} \sum_{l=1}^{s} I(X_k = x_i, Y_l = y_j)$.

H_0 : az ismérvek függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j} \quad \forall i, j$-re

H_1: az ismérvek nem függetlenek

Próbast.: $T_n(\mathbf{X}, \mathbf{Y}) = \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{i,j} - N_{i\bullet} N_{\bullet j}/n)^2}{N_{i\bullet} N_{\bullet j}/n} \right) \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{(r-1)(s-1)}$

elo.-ban

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

Ha $r = s = 2$, akkor a próbastatisztika $T_n = n \cdot \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\bullet}N_{2\bullet}N_{\bullet 1}N_{\bullet 2}}$-re egyszerűsödik, az aszimptotikus eloszlás pedig 1 szabadságfokú χ^2.

Feladat

E**12.)** Egy webtervező azt gyanítja, hogy az általa létrehozott internetes vásárlás honlapján a vásárlások mértéke összefügg azzal, hogy milyen nap van a héten. Ennek a sejtésnek az ellenőrzésére egy héten keresztül adatokat gyűjt – összesen 3758 látogatót számlált meg:

Vásárlás	H	K	Sz	Cs	P	Sz	V	Össz.
Nem vásárolt	399	261	284	263	393	531	502	2633
1 vásárlás	119	72	97	51	143	145	150	777
Több vásárlás	39	50	20	15	41	97	86	348
Összesen	557	383	401	329	577	773	738	3758

Alkalmas statisztikai próbával döntsünk arról, hogy helyes-e a webtervező sejtése!