3. előadás - Szórásanalízis

Sztochasztikus rendszerek matematikája

2016. szeptember 19.

Alapfeladat

A vizsgálandó kérdések típusai:

- Szabályos-e egy dobókocka? - illeszkedés-vizsgálat
- Két minta azonos eloszlású-e? - homogenitás-vizsgálat
- Független-e egymástól két ismérv, pl. képesség, avagy vásárlási szokások, stb... - függetlenség-vizsgálat

A fenti feladatokra mind használható úgynevezett χ^2-próba. Az elnevezés a próbastatisztikák határeloszlására utal.

Illeszkedésvizsgálat

Adott $\xi_1, \ldots, \xi_n$ i.i.d. mintáról szeretnénk eldönteni, hogy egy adott, $x_1, \ldots, x_k$ értékű eloszlásból származik-e. Azaz

$$\mathrm{H}_0 : P(\xi_1 = x_j) = p_j,\ 1 \le j \le k, \quad \mathrm{H}_1 : \mathrm{H}_0 \text{ nem teljesül.}$$

A próbastatisztika

$$T := \sum_{j=1}^{k} \frac{(np_j - \nu_j)^2}{np_j} \sim \chi^2_{k-1}$$

ahol $\nu_j = \sharp\{i : \xi_i = x_j\}$, a mintában előforduló x_j értékek száma. Adott α mellett a kritikus értéket a

$$P(T > a) = \alpha$$

összefüggés megoldása adja.

Homogenitásvizsgálat

Adott $\xi_1, \ldots, \xi_m$ és $\eta_1, \ldots, \eta_n$ független minták esetén azt szeretnénk eldönteni, hogy a két minta eloszlása megegyezik-e.

Legyen $(x_1, \ldots, x_k)$ az értékhalmaz, ν_j az x_j érték gyakorisága a ξ minta esetén, (μ_j) az η minta esetén, (p_j) és (r_j) pedig az eloszlások, $j = 1, \ldots, k$. Ekkor

$$\mathrm{H}_0 : p_j = r_j,\ 1 \le j \le k, \quad \mathrm{H}_1 : \mathrm{H}_0 \text{ nem teljesül.}$$

A próbastatisztika

$$T := \frac{1}{mn} \sum_{j=1}^{k} \frac{(n\mu_j - m\nu_j)^2}{\nu_j + \mu_j} \sim \chi^2_{k-1}$$

Adott α mellett a kritikus értéket a $P(T > a) = \alpha$ összefüggés megoldása adja.

Függetlenségvizsgálat

Adott két szempont és n megfigyelés, az első szempont szerint k, a második szerint pedig l osztály. Független lesz-e a két szempont egymástól? Legyen

$$A_i = \{\text{az első szempont szerint az } i.\text{ kategóriába esik a megfigyelés}\},$$

$$B_j = \{\text{a második szempont szerint a } j.\text{ kategóriába esik a megfigyelés}\}.$$

Ekkor

$$\mathrm{H}_0 : P(A_i \cap B_j) = P(A_i)P(B_j)\,\forall i, j, \quad \mathrm{H}_1 : \mathrm{H}_0 \text{ nem teljesül.}$$

A próbastatisztika itt is χ^2 eloszlású $(k-1)(l-1)$ szabadságfokkal. Adott α mellett a kritikus értéket a $P(T > a) = \alpha$ összefüggés megoldása adja.

Feladat

Tekintsük a gyakorlaton hallott példát három adatsorral!

Brand A	Brand B	Brand C
46	46	43
42	39	31
45	43	40
30	29	36
43	44	36
41	46	45
35	33	38
43		36
34		
41		
10 db	7 db	8 db

Feladat: Teszteljük a három féle abroncs élettartamának egyezését!

Két minta esetén alkalmazhatnánk kétmintás t-próbát, de most 3 csoportunk van. Ilyen esetekben használható a szórásanalízis vagy ANOVA (ANalysis Of VAriance).

Alapvető fajtái:

- Egyszempontos ANOVA: egyetlen faktor/magyarázó változó esetén
- Többszempontos ANOVA: több faktor esetén kereszthatásokkal is számolva

Többféle általánosítás vagy kiterjesztett modell is létezik, például mixed ANOVA vagy repeated measures ANOVA, stb.

Alapötlet

Alapötlet: a mintából számolt összvarianciát 2 részre bontjuk:

- csoportokon belüli variancia, (MS_{within})
- csoportok közötti variancia, $(MS_{between})$

Ezeket hasonlítjuk össze F-próbával.

Varianciákat (szórásnégyzeteket) vizsgálunk, mégis átlagokra hozunk döntést!
A kétmintás t-próba ennek speciális esete, és ugyanazt az eredményt is adja ugyanarra a mintára.

Egyszempontos ANOVA

A hipotézisek:

$$\mathrm{H}_0 : \mu_1 = \ldots = \mu_k = \mu, \quad \mathrm{H}_1 : \exists i : \mu_i \neq \mu$$

Kiinduló adatok:

	1. csop.	2. csop.	...	k. csop.
	x_{11}	x_{12}	...	x_{1k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$x_{N_1 1}$	$x_{N_2 2}$	...	$x_{N_k k}$
Elemszám	N_1	N_2	...	N_k
Átlag	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$

A teljes minta elemszáma: $N = N_1 + \ldots + N_k$, a teljes minta átlaga pedig $\bar{x}$.

Egyszempontos ANOVA

A minta teljes varianciája:

$$SS^2 = \sum_{j=1}^{k} \sum_{i=1}^{N_j} (x_{ij} - \bar{x})^2.$$

Mivel

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}) \quad \text{és} \quad N - 1 = (N - k) + (k - 1),$$

így a csoporton belüli és a csoportok közötti variancia

$$MS^2_{within} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{N_j} (x_{ij} - \bar{x}_j)^2}{N - k}$$

és

$$MS^2_{between} = \frac{\sum_{j=1}^{k} N_j (\bar{x}_j - \bar{x})^2}{k - 1}.$$

Egyszempontos ANOVA

A két variancia összehasonlítására egyoldali F-próbát alkalmazunk

$$F = \frac{MS^2_{within}}{MS^2_{between}}$$

alakban, mert csak az érdekel bennünket, hogy a belső szórás nagyobb-e, mint a minták közti szórás.

Feltételek: függetlenség, normalitás, homogenitás.

Tulajdonságok: robusztus a teszt, azaz nem túl érzékeny a feltételek sérülésére.

A nullhipotézis elutasítása esetén ún. Post Hoc teszteket kell alkalmazni annak eldöntésére, hogy a szignifikáns eltérés a minták mely tagjai közt lépnek fel.

Egytényezős varianciaanalízis

ÖSSZESÍTÉS

Csoportok	*Darabszám*	*Összeg*	*Átlag*	*Variancia*
Brand A	10	400	40	27,33333
Brand B	7	280	40	44,66667
Brand C	8	305	38,125	19,83929

VARIANCIAANALÍZIS

Tényezők	*SS*	*df*	*MS*	*F*	*p-érték*	*F krit.*
Csoportok között	19,125	2	9,5625	0,322229	0,727894	3,443357
Csoporton belül	652,875	22	29,67614			
Összesen	672	24				