Matematikai statisztika

Informatika alapszak, "B" 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

5. stat. előadás

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

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

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valamilyen eloszláscsaládból származik, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbastatisztika: $T_n(\mathrm{X}) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ 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

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

H_1: a nullhipotézis tagadása

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

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

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: $1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}$

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

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

Függetlenségvizsgálat

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

$p_{i,j} = P(\text{egy megfigyelés az } (i,j) \text{ osztályba kerül})$
$N_{i,j} =$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 $N_{i,j} = \sum_{k=1}^{n} I(X_k = x_i, Y_k = y_j)$.

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

elo.-ban

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

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.

Regresszióelemzés, lineáris modell – motiváció

- Gyakorlati szempontból a félév egyik legfontosabb témája!
- Az 1. órai kérdőíves felmérés alapján mennyire magyarázható jól
  - a hallgatók testmagassága a súlyuk segítségével?
  - a hallgatók testmagassága a súlyuk és a cipőméretük segítségével?
- Egy részvény holnapi árfolyamát hogyan jelezzük előre a tegnapi, tegnapelőtti, stb. árfolyamadatok segítségével?
- Egy gazda földvásárlási dilemmája – egy bizonyos földterületen a várható termésátlag mennyire jelezhető előre a földterület fontosabb jellemzői alapján (a talaj kémhatása, a $CaCO_3$ megjelenési mélysége, a humusztartalom, topográfiai helyzet)?
- Meg lehet-e becsülni annak az esélyét, hogy valaki élete során megbetegszik tüdőrákban? Hogyan modellezzük ezt? Például: megbetegedés esélye $\longleftarrow$ dohányzik-e, hány éven át dohányzott élete során, van-e tüdőrákos a közeli rokonságban, van-e egyéb tüdőbetegsége, poros/füstös helyen dolgozik-e?

Regresszióelemzés – bevezetés

Legyenek Y, $X_1, \ldots, X_p$ véges szórású valószínűségi változók, amik egy véletlen jelenség egy-egy jellemzői.

A regresszióelemzés célja: a bennünket különösen érdeklő Y valószínűségi változó "minél jobb" közelítése az $X_1, \ldots, X_p$ valószínűségi változók segítségével.

Y elnevezései: eredményváltozó, függő változó, endogén változó
X_i-k elnevezései: magyarázó változók, független változók, exogén változók

Általában megfigyeléseink vannak, amik az $(Y, X_1, \ldots, X_p)^T$ valószínűségi vektorváltozó realizációinak tekinthetők:
$(y_i, x_{i,1}, \ldots, x_{i,p})^T \quad i = 1, 2, \ldots, n \qquad$ általában $n >> p$
Feltehetjük, hogy az y_i megfigyelések rendszerint mérési eredmények, amik sajnos pontatlanok. A mérési hibát ε_i-vel fogjuk jelölni, amiről természetes feltétel, hogy legyen 0 várható értékű és véges σ szórású valószínűségi változó.

Regresszióelemzés

Legyenek Y, X, $X_1, \ldots, X_p$ véges szórású valószínűségi változók,
$c, a, b_1, \ldots, b_p$ valós számok.
Jelölje $\mathrm{X} = (X_1, \ldots, X_p)^T$, $\mathrm{b} = (b_1, \ldots, b_p)^T$ vektorokat.

	Feladat	Megoldás
a.)	$\min_c E(Y-c)^2$	$\hat{c} = EY$
b.)	$\min_{f:\mathbb{R}\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X))^2$	$\widehat{f}(X) = E(Y\|X)$
c.)	$\min_{a,b} E(Y-(a+bX))^2$	$\hat{b} = \frac{cov(X,Y)}{D^2X}$, $\hat{a} = EY - \hat{b}EX$
d.)	$\min_{f:\mathbb{R}^p\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X_1,\ldots,X_p))^2$	$\widehat{f}(X_1,\ldots,X_p) = E(Y\|X_1,\ldots,X_p)$
e.)	$\min_{a,b_1,\ldots,b_p} E\left(Y-(a+\sum_{i=1}^{p} b_i X_i)\right)^2$ [Többváltozós lineáris regresszió]	$\hat{\mathrm{b}} = (cov(\mathrm{X},\mathrm{X}))^{-1}\, cov(\mathrm{X}, Y)$ $\hat{a} = EY - \sum_{i=1}^{p} \hat{b}_i EX_i$

$E(Y|X)$: feltételes várható érték

Lineáris modell

- A modell: $\mathrm{y} = X\mathrm{b} + \varepsilon$
- $F := \mathrm{Im}X \quad \rightsquigarrow X$ képtere
- $r :=$rang(X), általánosan $r \leq p$, teljes rangú esetben $r = p$
- Paraméterbecslés: $\widehat{\mathrm{b}} = (X^T X)^{-1} X^T \mathrm{y}$
- Projekció az F altérre: $P_F = X(X^T X)^{-1} X^T$
- Becsült értékek: $\widehat{\mathrm{y}} := X\widehat{\mathrm{b}}$
- Reziduálisok: $\widehat{\varepsilon} = \mathrm{y} - \widehat{\mathrm{y}}$
- Reziduális négyzetösszeg: $\mathrm{RNÖ} := \mathrm{SSR} \|\widehat{\varepsilon}\|^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$
- Teljes négyzetösszeg: $\mathrm{NÖ}(= \mathrm{SS}) = \sum_{i=1}^{n} (y_i - \overline{y})^2$
- Determinációs együttható: $R^2 = 1 - \frac{\mathrm{RNÖ}}{\mathrm{NÖ}} = \frac{\mathrm{NÖ} - \mathrm{RNÖ}}{\mathrm{NÖ}} \quad \rightsquigarrow$ az eredményváltozó változékonyságának hány %-át magyarázza a regressziós modell
  Értéke 0 és 1 között lehet. Minél nagyobb, annál jobb.

Lineáris modell

- Korrigált determinációs együttható: $R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}\frac{SSR}{SS}$ $\rightsquigarrow$ egy lehetséges modellválasztási kritérium, minél nagyobb, annál jobb
- Akaike-féle információs kritérium: $AIC = 2(p+1) - 2\log \widehat{L}$, ahol $\widehat{L}$ a likelihood-függvény értéke akkor, ha az ML-becslést használjuk (normális eloszlású hibáknál ez megegyezik a legkisebb négyzetes becsléssel)
  Ez is egy lehetséges modellválasztási kritérium, minél kisebb, annál jobb.

Hipotézisvizsgálat a lineáris modellben

- t-próba az egyes együtthatókra (feltételezzük a hibák normális eloszlását): $H_0 : b = 0$, $H_1 : b \neq 0$
- A próbastatisztika: $t = \frac{\hat{b}}{D(\hat{b})}$, ez $n-1$ szabadságfokú t-eloszlású, ha igaz a H_0.
- Az egy magyarázó változós esetben ($y \sim a + bx$) $D^2(b) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$ és $D^2(a) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$.
- A reziduálisok: $r_i = y_i - \hat{y}_i$, ebből becsülhető a hiba szórásnégyzet: $\hat{\sigma}^2 = \sum_{i=1}^{n} r_i^2/(n-p)$
- A szokott módon konfidencia intervallum is konstruálható a becslésekhez

A regressziós modell "felépítése"

Ha p magyarázó változónk van, akkor 2^p modell közül kell a legjobbat megkeresni. Több módszer közül lehet választani:

- Nagyról kicsire (hátulról előre): először az összes magyarázó változót bevesszük, majd egyenként a legkevésbé szignifikánsat kivesszük egészen addig, míg mindegyik szignifikáns lesz
- Kicsiről nagyra (elölről hátulra): egyesével azt vesszük hozzá, amelyikkel a legjobban illeszkedő modellt kapjuk a következő lépésben. Vége: ha bármelyik, még a modellen kívüli magyarázó változót bevéve, már nem javul a modell illeszkedése.