HORVÁTH JENŐNÉ DR.[1]

A Bayes-statisztika és alkalmazása

„Ha nem tudsz más megoldást,
vedd elő a bayesi módszert!"

Bevezetés

A klasszikus valószínűség-számítás általunk jól ismert BAYES-tételén alapuló, ma BAYES-statisztikának nevezett statisztikai megközelítés a XX. század közepén szinte megrengette a statisztika világát. Teljesen újszerű szemlélete sok vitát váltott ki a klasszikus statisztika követői között. A bírálók elsősorban a külső információk felhasználását, és a szubjektív valószínűségek megjelenését kifogásolták. Ma már a kutatásokban és a gyakorlatban is egyre népszerűbbé válik ez az új módszer. Ugyanakkor Magyarországon még mindig különlegességnek számít az alkalmazása. Előadásomban először a BAYES-statisztika lényegét szeretném ismertetni, majd annak egy változatát a hierarchikus BAYES-módszert mutatom be. Végül a fogyasztói preferenciák vizsgálata során felmerült olyan problémának a megoldását szeretném megmutatni a hierarchikus BAYES-módszer segítségével, amire a hagyományos statisztika nem alkalmazható.

[1] BGF Külkereskedelmi Főiskolai Kar Matematika–Statisztika Tanszék, főiskolai docens.

A BAYES-módszer lényege

A bayesi megközelítés alapja BAYES tétele, ami folytonos valószínűségi változókra a következőképpen fogalmazható meg:

$$f(\Theta|x) = \frac{f(x|\Theta) \cdot f(\Theta)}{f(x)}$$

ahol:
x: a megfigyelések véletlen vektora;
Θ: a valószínűségi változónak tartott paramétervektor;
$f(\Theta \mid x)$, $f(x \mid \Theta)$: feltételes sűrűségfüggvények.

Mindezt átírva:

$$f(\Theta|x) \propto f(x|\Theta) \cdot f(\Theta),$$

ahol:

$\frac{1}{f(x)}$: az arányossági tényező;

$\cdot f(\Theta)$: Θ a priori sűrűségfüggvénye, a modellünk paramétereire vonatkozó előzetes ismereteinket tartalmazza. Ez lehet szubjektív és objektív, vagyis korábban végzett elemzésekből származó, adatokon nyugvó információ vagy szubjektív feltételezésekből, megérzésekből származó, nem adatokon alapuló információ. Legtöbbször nem is különböztetjük meg őket.

$f(x|\Theta)$: a mintavételi statisztikából jól ismert likelihood függvény. Ez hordozza a mintából származó információkat. Azt mutatja meg, hogy adott eloszlás és különböző paraméterek esetén mennyire hihető (valószerű), hogy éppen a szóban forgó minta adódik a mintavétel során.

$f(\Theta|x)$: az a posteriori sűrűségfüggvény.

Ebből az alakból jól kiolvasható a bayesi felfogás lényege: az *a posteriori* eloszlás sűrűségfüggvénye arányos az *a priori* sűrűségfüggvény és a likelihood függvény szorzatával, vagyis az *a posteriori* számára a mintán kívüli előzetes információt az *a priori* sűrűségfüggvény, a mintából származó információt pedig a likelihood függvény közvetíti.

Az arányossági tényezőt pedig úgy kell megválasztani, hogy az

$$\int_x f(\Theta|x) = 1$$

legyen.

A klasszikus és a bayesi szemlélet közötti különbségek

A bayesi szemléletű statisztikus minden lehetséges, fellelhető információt felhasznál és ezeket kombinálja a mintavétel eredményével.

A bayesi statisztikában lehetővé válik a szubjektív vélemények egzakt kezelése a szubjektív valószínűségek használatával.

A bayesi becslésben a becsülni kívánt paraméter nem egy rögzített érték, hanem valószínűségi változó.

A bayesi statisztika tartalmazza speciális esetként (a külső információk teljes hiánya) a klasszikus elméletet. Ma már tudjuk, hogy a BAYES-tételnek ez az átfogalmazott formája mindenféle statisztikai modellnél alkalmazható, az összefüggés ugyanakkor igen egyszerűnek tűnik. Éppen ezen tulajdonságok (egyszerűség és az egységesség) miatt várható egyre szélesebb körű alkalmazása.

Hierarchikus BAYES-módszer

Tegyük fel, hogy a megfigyelések alapján ismerjük az $f(x|\theta)$ sűrűségfüggvényt, ami r db $\theta = (\theta_1, \theta_2,...\theta_r)$ ismeretlen paramétertől függ, amire vonatkozóan van egy $f(\theta)$ prior sűrűségfüggvényünk. Tegyük fel, hogy θ_i–k független azonos eloszlásúak és együttes eloszlásuk függ egy újabb paramétertől, amit hiperparaméternek hívunk. Ha ez a hiperparaméter nem ismert, egy újabb prior (hyperprior) ismeretre van szükségünk, ami kifejezi a paraméter lehetséges értékére vonatkozó hitünket. Ebben az esetben hierarchikus priorról beszélünk.

Hierarchikus BAYES-módszer alkalmazása a fogyasztói preferenciák vizsgálatánál

A conjoint analízis ma már jól ismert technika a több-attribútumú alternatívák közötti fogyasztói preferenciák vizsgálatára. Az eddig alkalmazott eljárások azonban feltételezik, hogy a megkérdezettek a profilok teljes halmazára válaszolnak. A tapasztalat azonban azt mutatja, hogy az emberek nem mindig képesek vagy nem hajlandók a profilok teljes halmazára reagálni. A felmerülő probléma megoldása lehetségessé válik a hierarchikus BAYES-módszer alkalmazásával.

Mi a *full profile* eljárást alkalmazzuk, vagyis a amikor a válaszadó a teljes termékről mond véleményt.

Az adott termék vagy szolgáltatás – a továbbiakban egységesen profil- értékelését meghatározó tulajdonságokat attribútumoknak vagy faktoroknak, konkrét értékeit pedig szinteknek nevezzük. A nagyszámú faktorkombináció helyett annak megfelelő részhalmazával dolgozunk *(„fractional factorial design").* Csak a főhatásokat vesszük figyelembe *(„orthogonal array"),* az egyes faktorokhoz „fontosságokat" *(„importance"),* azok szintjeihez „részhasznosságokat", részértékeket *(„utility")* rendelünk.

A preferenciák és a faktorszintek között lineáris kapcsolatot feltételezve véletlen hatású lineáris modellt alkalmazunk.

Az egyéni szintű részértékek és heterogenitásuk függ az egyénenkénti profilok számától és a vizsgálatban résztvevő személyek számától.

Tételezzük fel, hogy n személy válaszol a profilok halmazára és az i-edik személy J_i profilról mond véleményt. Induljunk ki a következő véletlen hatású lineáris modellből:

$$Y_i = X_i\beta_i + \varepsilon_i$$
$$\beta_i = \Theta z_i + \delta_i$$

Y_i : az i-edik személy metrikus válaszának a J_i dimenziós vektora

Xi : a tervmátrix

βi : az i-edik válaszadó részértékeit tartalmazó vektor

Az egyedi szintű részérték heterogenitást a második egyenlet írja le, ahol

θ: a hiperparaméter

Célunk a részértékek meghatározása, amiből a fogyasztói preferenciákra tudunk következtetni. A modellünk a β_i-t mint valószínűségi változót kezeli. Így csak a BAYES-becslés jöhet számításba. Mivel van egy hiperparaméterünk, a hierarchikus BAYES-módszert fogjuk alkalmazni.

A hierarchikus BAYES-módszer alkalmazása egy konkrét probléma megoldásában

Az „Egészséges Álom Tejüzem" menedzsmentje elhatározta, hogy új terméket jelentet meg a piacon. A conjoint analízis végrehajtására cégünket kérték fel. Az előzetes tájékozódás alapján az *1. táblázat* szerinti attribútumokat találtuk a vizsgálatra megfelelőnek.

1. táblázat
Vizsgálati szempontok

Jel	Megnevezés	–1	1
A	zsírtartalom	1,5 %	2,8 %
B	csomagolás	zacskós	dobozos
C	tömeg	0,5 liter	1 liter
D	tartósság	féltartós	friss
E	Energiatartalom 100 grammban	265 kJ	alacsonyabb
F	Fehérjetartalom 100 grammban	3,3 g	alacsonyabb
G	szénhidrát 100 grammban	4,6 g	alacsonyabb
H	kalcium 100 grammban	115 mg	alacsonyabb
I	ár	150 Ft/l	alacsonyabb
J	pasztőrözöttség	pasztőrözött	ultra pasztőrözött
K	Tartósítószert tartalmaz	igen	nem
L	ízesített	nem	igen

Ez összesen 4096 kombinációt jelentene, de az *orthogonal array* 16 megfelelően kiválasztott kombináció alapján megbízható eredményt nyújt (*2. táblázat*).

2a. táblázat
Profilok

Profil	A	B	C	D	E	F	G	H	I	J	K	L
1	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1
2	–1	1	–1	1	–1	–1	1	–1	1	1	–1	1
3	1	–1	1	–1	1	–1	1	1	–1	1	–1	1
4	1	1	1	1	1	–1	–1	1	1	–1	–1	–1
5	–1	1	1	–1	–1	1	–1	1	1	1	1	1
6	–1	–1	1	1	–1	1	1	1	–1	–1	1	–1
7	1	1	–1	–1	1	1	1	–1	1	–1	1	–1
8	1	–1	–1	1	1	1	–1	–1	–1	1	1	1
9	1	1	–1	–1	–1	1	–1	1	–1	–1	–1	1
10	1	–1	–1	1	–1	1	1	1	1	1	–1	–1
11	–1	1	1	–1	1	1	1	–1	–1	1	–1	–1
12	–1	–1	1	1	1	1	–1	–1	1	–1	–1	1
13	1	–1	1	–1	–1	–1	–1	–1	1	1	1	–1
14	1	1	1	1	–1	–1	1	–1	–1	–1	1	1
15	–1	–1	–1	–1	1	–1	1	1	1	–1	1	1
16	–1	1	–1	1	1	–1	–1	1	–1	1	1	–1

2b. táblázat
Ellenőrzési profilok

Profil	A	B	C	D	E	F	G	H	I	J	K	L
1	1	1	1	1	–1	1	1	–1	–1	1	1	1
2	1	1	–1	–1	1	1	–1	–1	1	1	–1	1
3	1	–1	–1	–1	–1	–1	1	1	1	–1	1	1
4	–1	–1	1	1	–1	1	–1	–1	1	1	–1	–1

A megkérdezettek száma: ..... 110 fő
Az egy megkérdezettre jutó profilok száma: 16
Az egy megkérdezettre jutó ellenőrzési profilok száma: 4

Az ellenőrzési profilok *(holdout)* megítélésre kerülnek, de a számítások során nem használjuk őket.

Tekintsük azt az esetet, amikor minden válaszoló a tervmátrixban szereplő összes profilt értékeli. A *3. táblázat* az egyesített minta (110 fő) aggregált eredményeit mutatja az OLS legkisebb négyzetek becslés alapján, ami a fogyasztói

heterogenitást nem veszi figyelembe (fix hatású lineáris modell), ezért az STD (standard eltérés) hibák a mintából történő becslésben levő bizonytalanságot mérik. A táblázatból az alábbi következtetések állapíthatók meg:

- A fogyasztók a következő tulajdonságokkal rendelkező tejet preferálják: 1,5% zsírtartalmú, zacskós, literes, féltartós, 260 kJ energiatartalmú, 3,3 grammnál alacsonyabb fehérjetartalmú, 4,6 g szénhidráttartalmú, 115 mg kalcium tartalmú, 150 Ft/l-nél alacsonyabb árú, tartósítószert nem tartalmazó, ízesített.
- A legfontosabb attribútumok: 150 Ft/liternél alacsonyabb ár, pasztőrözött legyen és 1,5% zsírtartalmú.
- A legkisebb mértékben preferált attribútumok: tömeg, kalciumtartalom, tartósítószer tartalom.

3. táblázat
Az egyesített minta aggregált eredményei
Estimated Coefficients

Variable	Coefficient	STD Error	T Value	Variable	Coefficient	STD Error	T Value
Intercept	0,4809	0,0061	78,9051	G	–0,0210	0,0061	–3,4425
A	–0,0381	0,0061	–6,2576	H	–0,0027	0,0061	–0,4502
B	–0,0175	0,0061	–2,8683	I	0,1219	0,0061	20,0015
C	0,0017	0,0061	0,2824	J	–0,0515	0,0061	–8,4482
D	–0,1100	0,0061	–1,8019	K	0,0093	0,0061	1,5316
E	–0,0161	0,0061	–2,6352	L	0,0191	0,0061	3,1349
F	0,0212	0,0061	3,4835				

R-Squared: 0,2445	Adjusted R Squared: 0,2393	Standard Error of the Estimate: 0,2557

A valódi preferenciák és a becsült preferenciák közötti korrelációs együttható a négy ellenőrző profilra vonatkozóan: 0,616, ami közepesnek mondható.

Az előzőekben nem vettük figyelembe a fogyasztói heterogenitást. A mi esetünkben azonban a válaszadók csak 14 profilt választottak a tervmátrixban szereplő 16 helyett. A kérdés az, hogy tudunk-e olyan módszert találni, amellyel a preferált termékhez eljuthatunk ebben az esetben is. A hierarchikus Bayes-becslés segítségével próbáljuk lefedni ezt a heterogenitást úgy, hogy két véletlenszerűen kiválasztott profillal csökkentjük az egy fogyasztóra jutó profilok számát az előző tervmátrixhoz képest.

A *4. táblázat* egyértelműen mutatja, hogy a preferenciák változatlanok maradtak.

Csökkentsük tovább kettesével a profilok számát és ismételjük addig, amíg összesen két profil marad. A *4. táblázat* eredményei azt mutatják, hogy a gyengén preferált attribútumok esetén a preferált attribútum szint is felcserélődött a profilok számának csökkenésével. A többinél csak a sorrendben történt váltás.

Ebben az esetben a standard eltérések az egyéni szintű becslések közötti megfigyelt standard eltéréseket jelentik (STD). Az egyéni szintű becslések szóródása jelentősen meghaladja az egyesített minta bizonytalanságát, amit az említett fix hatású modellben kaptunk. Az átlagokra és standard eltérésekre vonatkozó OLS és hierarchikus BAYES-becslés becslés a teljes tervmátrixra közel azonos, ha csökkentjük a profilok számát pedig közelítenek a teljes tervmátrix esetén kapotthoz.

A korreláció erőssége a kétféle becslés esetén, a hierarchikus BAYES-becslésnél csökkentve az egy fogyasztóra jutó profilok számát. A profilok csökkenésével a korrelációs együttható értéke is csökken. A profilok számát felére csökkentve még viszonylag elfogadható eredményeket kapunk a korrelációra. Az utolsó két táblázat azt mutatja, hogy a hierarchikus BAYES-becslés modell alkalmas lehet a részérték heterogenitás becslésére hat vagy több profil esetén.

Következtetés

A példa is mutatja, hogy a hierarchikus BAYES-módszerrel is lehet az OLS-hez hasonló szintű pontosságot elérni a részhasznosságokban (részértékekben), ha a fogyasztók különböző csoportjai a profilok különböző részhalmazaira válaszolnak és minden részhalmazba kevesebb profil tartozik, mint, amennyit a teljes tervmátrix tartalmaz. Mindez lehetővé teszi, hogy a marketing kutatók abban az esetben is viszonylag pontos eredményekhez jussanak a fogyasztói preferenciák vizsgálata esetén, ha a megkérdezettek nem akarnak vagy nem tudnak minden terméket értékelni. A további vizsgálódás útja igen széles lehetőségeket rejt magában. Érdemes lenne gyakorlati alkalmazásokkal még alátámasztani a hierarchikus BAYES-becslés conjoint analízis előnyeit, esetleg az apriori feltevéseken változtatva tovább lépni vagy más még megoldatlan problémáknál próbálkozni a BAYES-módszer alkalmazásával is, hátha segít.

Irodalom

CASELLA, G., GEORGE, E. I.: Explaining the Gibbs Sampler. American Statistical Association. 1992.

KŐRÖSI G., MÁTYÁS L., SZÉKELY I.: Gyakorlati ökonometria. KJK. Bp. 1990.

LENK, P. J., DESARBO, W. S., GREEN, P. E., YOUNG, M. R.: Hierarchical Bayes Conjoint Analysis: Recovery of Part-Worth Heterogeneity from Incomplete Designs in Conjoint Analysis. March 10, 1994.

LINDLEY, D. V., SMITH, A. F.: Bayes Estimates for the Linear Model. Journal of the Royal Statistical Society, Series B, 34, 1-41.

PETER, M. L.: Bayesian Statistics. Arnold, Great Britain, 1989.

SMITH, A. F.: A General Bayesian Linear model. Journal of the Royal Statistical Society, Series B, 35, 67-75..

TULL, D. S., HAWKINS, D. I.: Marketing Research. HUMBERSIDE. 1993.

4. táblázat
Heterogenitás vizsgálata

Pro-fils		Inter-cept	A	B	C	D	E	F	G	H	I	J	K	L	Error Var.
Ordinary Least Squares															
16	Mean	0,4810	-0,0381	-0,0175	0,0017	-0,0110	-0,0161	0,0212	-0,0210	-0,0027	0,1220	-0,0515	0,0093	0,0191	2,1037
	STD	0,1307	0,0560	0,0423	0,0366	0,0368	0,0507	0,0523	0,0543	0,0365	0,1329	0,0761	0,0393	0,0472	2,6400
Hierarchical Bayes															
16	Mean	0,4811	-0,0384	-0,0175	0,0017	-0,0111	-0,0159	0,0210	-0,0213	-0,0029	0,1224	-0,0515	0,0094	0,0196	1,6185
	STD	0,1264	0,0522	0,0388	0,0331	0,0335	0,0473	0,0466	0,0495	0,0326	0,1285	0,0718	0,0354	0,0431	0,8245
14	Mean	0,4838	-0,0371	-0,0181	0,0020	-0,0100	-0,0138	0,0236	-0,0246	-0,0029	0,1227	-0,0499	0,0112	0,0213	1,2577
	STD	0,1258	0,0556	0,0435	0,0369	0,0422	0,0522	0,0473	0,0535	0,0386	0,1326	0,0728	0,0437	0,0475	0,8462
12	Mean	0,4824	-0,0349	-0,0178	0,0014	-0,0121	-0,0170	0,0238	-0,0208	-0,0012	0,1229	-0,0464	0,0136	0,0178	0,7781
	STD	0,1245	0,0593	0,0488	0,0436	0,0494	0,0537	0,0478	0,0581	0,0404	0,1263	0,0699	0,0475	0,0530	1,1359
10	Mean	0,4810	-0,0368	-0,0194	-0,0020	-0,0111	-0,0162	0,0208	-0,0154	0,0037	0,1230	-0,0509	0,0106	0,0192	0,1144
	STD	0,1121	0,0608	0,0463	0,0416	0,0500	0,0581	0,0551	0,0637	0,0450	0,1181	0,0648	0,0491	0,0592	0,0459
8	Mean	0,4867	-0,0364	-0,0243	-0,0131	-0,0136	-0,0105	0,0270	-0,0264	0,0082	0,1281	-0,0393	0,0024	0,0190	0,0531
	STD	0,0914	0,0494	0,0384	0,0408	0,0441	0,0558	0,0480	0,0542	0,0394	0,0962	0,0590	0,0459	0,0473	0,0041
6	Mean	0,4871	-0,0445	-0,0223	-0,0126	-0,0140	-0,0225	0,0199	-0,0193	0,0076	0,1297	-0,0352	-0,0004	0,0170	0,0598
	STD	0,0708	0,0432	0,0368	0,0402	0,0437	0,0507	0,0392	0,0453	0,0363	0,0778	0,0533	0,0423	0,0443	0,0024
4	Mean	0,4817	-0,0345	-0,0318	-0,0126	-0,0177	-0,0144	0,0261	-0,0230	-0,0008	0,1265	-0,0362	-0,0153	0,0226	0,2761
	STD	0,0456	0,0384	0,0315	0,0375	0,0388	0,0364	0,0318	0,0367	0,0311	0,0557	0,0464	0,0387	0,0376	0,0032
2	Mean	0,5142	-0,0328	-0,0176	-0,0138	-0,0157	-0,0221	0,0213	-0,0493	0,0196	0,1263	-0,0353	-0,0126	0,0259	0,4356
	STD	0,0275	0,0268	0,0252	0,0281	0,0252	0,0275	0,0254	0,0288	0,0260	0,0323	0,0296	0,0301	0,0268	0,0054