# DATUEN ANALISIA ETA ENUNTZIATUEN LOGIKA *

Yosu YURRAMENDI

## 1. Sarrera

Ikertzaike askoren lanak datuen jasoketa batekin hasten dira sarritan. Datu pila baten aurrean ez dago ikertzailearentzako, dakigunez, inolako araubiderik horren gainean gorputz teoriko bat eraikitzeko. Dena dela, zeregin horretan saiatzea ez da lan antzua, datuak nolabait deskribatu, mamitu eta sailkatu behar bait dira.

Bada teknika multzo bat zeregin hoietarako lehenengo urratsak ematen laguntzeko balio duena: estatistika deskribatzailea deritzana. Hona hemen John Tukey estatistikari ospetsuak zer zioen aspalditxo "Annals of Mathematical Statistics, vol. 33, 1962" aldizkarian "The Future of Data Analysis" izenburu pean (2. orr.):

> *"For a long time I have thouht I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt".*

Eta aurreraxeago:

> *"I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways for planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data".*

Beharbada, lerroaldi hauetan oso garbi geratu ez den zerbait, datu pila bat analisatzea zertan datzan jakitea da. Gehienetan, datuak zenbait objeturi behatutako ezaugarrietatik datoz; datuak analisatzea beraz, objetuen arteko erlazioak, ezaugarrien artekoak eta, objetu eta ezaugarrien artekoak azaltzean edo argitaratzean datza.

Datuen analisian erabili behar diren prozedura eta tekniken nolakotasunari honela deritzo John Tukey-k aipatutako (6. orr.):

> *"Data analysis, and the parts of statistics which adhere*

* El abstract inglés y la versión castellana de este artículo siguen en las páginas 275-291.

THEORIA - Segunda Época,
Año I (1985), Nº 1, pp. 259-273.

*to it, must then take on the characteristics of a science rather those of mathematics, specifically:*

*(b-1) Data analysis must seek for scope and usefulness rather than security.*

*(b-2) Data analysis must be willing to err moderately often in order that inadecuate evidence shall more often suggest the right answer.*

*(b-3) Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proof or stamps of validity".*

Eta geroxeago:

*"All sciences have much of art in their makeup. (...) As well as teaching facts and well-established structures, all sciences must teach their apprentices how to think about things in the manner of that particular science, and wath are its current beliefs and practices. Data analysis must do the same."*

Hemen "structure" hitza axioma eta definizio multzo bat bezala ulertzen dugu. Arrazoinamendu -(logiko)- aren bidez bereizgarriak edo propietateak ondorioztatzen dira.

Idazki honen garaietatik gaurko egunetara kalkulugailuen garapen eta hedapena itzelak izan dira. Beren kalkuluetarako gaitasun eta oroimenak, lehen bururatu ere ezin ziren eginkizunak ekarri ditu, eta honela, aztergarriak diren datu pilak gero eta haundiagoak dira.

Ikus dezagun orain, datuen analisiaren hastapenak zeintzu diren hamar bat urte beranduago, eskola frantzesako burua edo izan den Jean-Paul Benzecri-ren iritziz, bere inguruko lankide batzurekin argitaratutako "L'Analyse des Données. Tome I: La taxinomie. Tome 2: L'Analyse des correspondances. Edt. Dunod, 1973, 4ème édition 1982" idazlanean (3-17 orr.):

*"1ér Principe: Statistique n'est pas probabilité. Sous le nom de statistique mathématique, des auteurs (qui, je vous le dis en français, n'écrivent guère dans notre langue ...) ont edifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes*

*typologiques".*

.....

*"2ème Principe: Le modèle doit suivre les données, non l'inverse. (...) Mais ce dont nous avons besoin c'est d'une méthode rigoureuse qui extraie des structures à partir des données.*

.....

*3ème Principe: Il convient de traiter simultanément des informations concernant le plus grand nombre possible de dimensions.*

.....

*4ème Principe: Pour l'analyse des faits complexes et notamment de faits sociaux, l'ordinateur est indispensable. Principe évidemment vrai ... mais qu'en eussent pensé nos pères les Gaulois il y a 15 ans?*

.....

*5ème Principe: Utiliser un ordinateur implique d'abandonner toutes techniques conçues avant l'avènement du calcul automatique. Je dis technique, non science: (...)"*

Ikusten denez, bi irakasleek badute zenbait puntu elkarren artean, beren aurkezteko estiloak zeharo desberdinak izan arren. Beren ustez, badirudi, datuen analisia estatistika deskribatzaile klasikoaren itxura berria dela. Gure idazki honetan, hastapen hauei erantzuten dien eredu baten eraiketa azalduko dugu, guztiz orokorra dena eta hain zuzen enuntziatuen logikarekin lotura duena. Bertan zenbait arazo planteatuko dugu.

2. Galdeketaren eredua

Datu pila baten aurrean eredu bat eraikitzerakoan, ikuspegiren batetik edo bestetik, jasotako datuen berreraiketa baten ahalmena eman behar da. Gehienetan ordea, asmo hau nekez bete daiteke, datuen kodaketak berak informazio guztiaren zati bat besterik ez bait du biltzen.

Hemen azalduko dugun ereduak galdeketa izena du, gizazientzia delakoetan maiz erabili ohi dena gogoratzen du eta.

Bedi $\Omega$ aztergai diren objetuen multzoa. Aurrerantzean finitua dela joko dugu:

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

Bedi X, $\Omega$ multzoaren ezaugarri baten agertzeko eren multzoa,

edo konparazioari jarraika, $\Omega$-koek galdera bati eman liezizkioketen erantzunen multzoa. Hau ere finitutzat hartuko dugu:

$$X = \{x_1, x_2, \ldots, x_p\}$$

Bedi X*, $\Omega$-koei egindako galdera bat:

$$X^*: \Omega \to X$$

hau da, $\Omega$-ren aplikazio edo aldagai bat.

Bitez

$$\{X_n \mid n=1, \ldots, r\} \quad \text{eta} \quad \{X^*_n \mid n=1, \ldots, r\}$$

r galderari eman lekizkieken erantzunen multzo-sorta bat eta honi dagokion aplikazio-sorta. Honela, ereduak suposatzen duena zera da: $\Omega$ -ko bakoitzak galdera bakoitzari erantzuten diola, eta gainera, era bakar batez.

Bedi

$$\{e^*_{nk} \mid n=1, \ldots, r \quad y \quad k=1, \ldots, p_n\}$$

$\Omega$ -ren aplikazio-sorta bat $\{0,1\}$ multzoan, era honetan definituta:

edozein $\omega \in \Omega$, edozein n = 1, ..., r eta edozein k = 1, ..., $p_n$ -entzako baldin

$$X^*_n(w)=X_{nk},$$

bada orduan

$$e^*_{nk}(w)=1,$$

eta bestela

$$e^*_{nk}(w)=0.$$

Ohar daitekenez, multzo hau erantzun guztien identifikatzaileena da, funtzio karakteristikoena alegia. Multzo honen bereizgarri edo propietate bat hauxe da:

edozein $\omega \in \Omega$, eta edozein n = 1, ..., r -entzako

$$\sum_{k=1}^{p_n} e^*_{nk}(w) = 1$$

E ikurraz izendadutako dugu multzo hau.

Noski,

$$\bigcup_{n=1}^{r} X_n$$

eta E multzoen artean korrespondentzia biuniboko bat dago.

Eredu hau eraikitzerakoan, asmoa, esan bezala, $\Omega$ -koen artean izan daitezken erlazio nabarmenak (zehaztu beharko den zentzu batetan) argitaratzea da, hala nola erantzunen artekoak eta, objetu eta erantzunen artekoak ere, guztiok, erantzunei dagozkien identifikatzaileen bitartez.

Bedi

$$E^*: \Omega \to \{0, 1\}^p, \qquad p = \sum_{n=1}^{r} p_n,$$

izanik, era honetan definituta:
edozein $\omega \in \Omega$-rentzako

$$E^*(w) = (e^*_{nk}(w) \mid n=1, \ldots, r;\ k=1, \ldots, p_n).$$

Eraikitako eredua murriztu egingo dugu galdera edo aldagai dikotomikoen kasuetara, idazki honen xederako nahikoa izango bait da. Edozein n = 1, ..., r- rentzako $X_n$ multzoa bi elementuz osatuta dago:

$$X_n = \{x_n^+, x_n^-\},$$

eta bi hauei dagozkien identifikatzaileak $e^{*+}$ eta $e^{*-}$ dira. Arestian aipatutako bereizgarria beste honetan bihurtu da:
edozein $\omega \in \Omega$ eta edozein n = 1, ..., r -entzako

$$e^*_n{}^-(w) = 1 - e^*_n{}^+(w).$$

Hortaz, datorrenerako E * aplikazioaren pareko hau hartuko dugu gogotan:

$$E^{*+}: \Omega \to \{0, 1\}^r,$$

edozein $\omega \in \Omega$ -rentzako

$$E^{*+}(w) = (e^*_n{}^+(w) \mid n=1, \ldots, r).$$

## 3. Enuntziatuen logika

$X_n$, n = 1, ...., r multzoak definitzean, esan bezala, r galderari eman lekizkieken erantzunak definitzen ditugu. Hauek bi baldin badira, erraz pentsa liteke kualitate beten jabe izatea adierazi dezakela batek, eta ez izatea besteak.

Enuntziatuen logikan (logikaren mailarik apalena bestalde) bi enuntziatuen $X_k$, $X_l$ arteko loturak edo erlazioak, $(X_k \times X_l)$ multzoaren azpimultzoak dira. $(X_k \times X_l)$ bidekarketa cartesiarra 4 elementuz osatuta dago, eta dakigunez, $2^4$ lotura edo erlazio desberdin definitu litezke.

Arruntenak, adibidez, hauexek dira.

a) Konjuntzioa: $X_k \& X_l = \{(x_k^+, x_l^+)\}$

b) Disjuntzioa: $X_k v X_l = \{(x_k^+, x_l^+), (x_k^+, x_l^-), (x_k^-, x_l^+)\}$
$= \{(x_k^-, x_l^-)\}^c$

c) Baldintza: $X_k \to X_l = \{(x_k^+, x_l^+), (x_k^-, x_l^+), (x_k^-, x_l^-)\}$
$= \{(x_k^+, x_l^-)\}^c$

d) Baldintza bikoitza: $X_k \leftrightarrow X_l = X_k \to X_l \cap X_l \to X_k$
$= \{(x_k^+, x_l^+), (x_k^-, x_l^-)\}$

e) Bateraezinezko disjuntzioa[1]: $X_k \not\leftrightarrow X_l = \{(x_k^+, x_l^-), (x_k^-, x_l^+)\}$

Bestalde, $\{0,1\}^2$ -ren aplikazioak $\{0,1\}$ multzoan $2^4$ dira ere, eta beraz, korrespondentzia biuniboko bat egokitu daiteke erlazio eta aplikazioen artean parekaketa hau egin ondoren:

$(x_k^+, x_l^+)$ eta (1, 1);

$(x_k^+, x_l^-)$ eta (1, 0);

$(x_k^-, x_l^+)$ eta (0, 1):

$(x_k^-, x_l^-)$ eta (0, 0).

Adibideetan azaldutako erlazioak honela adierazi litezke deritzaten funtzio logikoren bitartez

Edozein $(e_1, e_2) \in \{0,1\}^2$ -rentzako

a) $f_{\&}(e_1, e_2) = e_1 e_2$

b) $f_{\vee}(e_1, e_2) = 1-(1-e_1)(1-e_2)$

c) $f_{\rightarrow}(e_1, e_2) = 1 - e_1(1-e_2)$

d) $f_{\leftrightarrow}(e_1, e_2) = 1-(e_1-e_2)^2$

e) $f_{\not\leftrightarrow}(e_1, e_2) = (e_1-e_2)^2$

$\prod_{n=1}^{r} X_n$ biderkaketa cartesiarraren azpimultzoak r enuntziatuen arteko erlazioak adierazten dituzte, eta horietako batzuek badute zentzurik hizkuntza mailan. Aurreko adibideetako erlazioak, esaterako, multzo berri honetan daude murgilduta:

a) $X_k \& X_l = \{(x_1, \ldots, x_r) | x_n \in X_n \ (n=1, \ldots, r) \ \& \ x_k = x_k^+ \text{ y } x_l = x_l^+\}$

b) $X_k \vee X_l = \{(x_1, \ldots, x_r) | x_n \in X_n \ (n=1, \ldots, r) \ \& \ x_k = x_k^+ \text{ o } x_l = x_l^+\}$

c) $X_k \rightarrow X_l = \{(x_1, \ldots, x_r) | x_n \in X_n \ (n=1, \ldots, r) \ \& \ x_k = x_k^- \text{ o } x_l = x_l^+\}$

d) $X_k \leftrightarrow X_l = (X_k \rightarrow X_l) \cap (X_l \rightarrow X_k)$

e) $X_k \not\leftrightarrow X_l = (X_k \leftrightarrow X_l)^c$

Bi dimentsiotan egin dugunaren antzeko parekaketa baldin badagokie $\{0,1\}^r$ -ko elementuei, adibideetako erlazioak funtzio logiko hauen bitartez adierazi litezke:

edozein $(e_n \mid n = 1, \ldots, r) \in \{0,1\}^r$ -rentzako

a) $f(e_1, \ldots, e_r) = e_k e_l$

b) $f(e_1, \ldots, e_r) = 1-(1-e_k)(1-e_l)$

c) $f(e_1, \ldots, e_r) = 1-e_k(1-e_l)$

d) $f(e_1, \ldots, e_r) = 1-(e_k-e_l)^2$

e) $f(e_1, \ldots, e_r) = (e_k-e_l)^2$

Logikaren ikuspegitik arazo bat enuntziatuen mailan planteatzen denean, beti aurkitu daiteke f funtzio logiko bat arazo horren ispilu dena (definitutako korrespondentzia biunibokoaren zentzuan).

Soluzioa $\{0,1\}^r$ -ren $f^{-1}(1)$ azpimultzoa da. Dakigunez, $f^{-1}(1) = \emptyset$ baldin bada erlazioari kontresana deitzen diogu; aitzitik, $f^{-1}(1) = \{0,1\}^r$ bada, tautologia.

Datuen analisiaren ikuspegitik ordea, arazoa bestelakoa da. Erantzunen identifikatzaileen bitartez, $\{0,1\}^r$ -ren azpimultzo batetara heltzen gara, eta arazoa zein funtzio logikoen soluzioa den jakitean datza, hala nola azpimultzo horren parte berezi batzuri zeintzu dagozkie (berezitasun hori aurrerago geratuko da zehaztuta).

$\Omega$-ren tamaina (kardinala) nahiko haundia denean (demagun N = 500) eta ezaugarrien kopurua ere (r = 30), $\Omega$-ren irudia $E^{*+}$aplikazioaren bitartez, aski haundia izaten da bi edota hiru ezaugarrien arteko erlazioak ( $(X_k \times X_l)$ eta antzekoak ez besteak) errazki azaltzeko (hura hauen soluzioa izateko alegia), nahiz $2^{30}$-ren ondoan ($\{0,1\}^{30}$ -ren kardinala) oso txikia izan; datuen ulerkuntzak bestalde, erlazio xinpleak eskatzen ditu (esaterako, konjuntzioak, disjuntzioak, gai gutxietako baldintzak, eta abar).

Aurkako bi oharpen hauen artean soluziobide bat bilatu beharra dago. Estatistikaren hildotik abiatuz hurbilketa bat egin ohi da: erlazio xinpleak bilatzen dira $\Omega$ edota bere parte berezi batzuren gehiengoarentzako ("gehiengo" hitza zehaztu beharko den zentzu batetan).

Zer nolako loturak aztertu nahi izateak metodoaren muinean egon behar du. Adibidez, bi galderaren arteko baldintzazko erlazioak aztertu nahi bagenitu, ez litzateke okerra izango

$$\sum_{w \in \Omega} (1-e^{*}_{k}{}^{+}(w)(1-e^{*}_{k}{}^{+}(w)))$$

bezalako zenbakiak kalkulatzea galdera bikote bakoitzarentzako ($\Omega$ osoan beteko balitz erlazio hori, batura horrek N balioko luke), eta haundienak hartzea; esaterako, datuei Guttman-en eskala [2] egokitzea ohizko praktika izan da orain dela urte batzu arte.

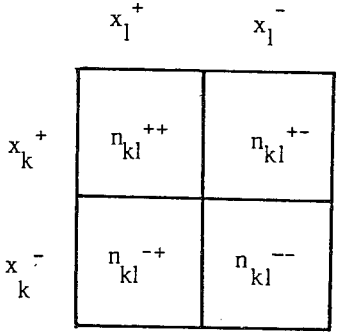
Beste adibide bat, bi galderaren arteko konjuntziozko erlazioena da. Bi galderaren elkartasun maila, biei batera baiezkoa eman dutenen kopuruan oinarritu liteke:

$$n_{kl}^{++} = \sum_{w \in \Omega} e^{*}_{k}{}^{+}(w) e^{*}_{l}{}^{+}(w)$$

Sarritan ordea, ez da sendoa izaten ezezkoei baliorik ez ematearen arrazoia: bietako galderaren bat alderantziz planteatuz gero, elkartasun maila desberdina izango zen. Halako kasu baten aurrean jokabiderik zuzenena izaten da galderaren bi erantzunak gogotan hartzea. Horrela bi galderari dagozkien lau gertakizunek, $X_k \times X_l$ -koek alegia, ezagubide berdina izaten dute. Bestalde, ez da informazio gehiagorik galtzen.

Dagikunez:

$$n_{kl}^{++} + n_{kl}^{+-} + n_{kl}^{-+} + n_{kl}^{--} = N$$




Datuen analisiaren metodoek indartsuak izan behar dute erlazioak aditzera emateko ikertzaleari: ezaugarrien azterketa orokor batetara bideratu behar dira, eta agerian utzi behar dute beren arteko loturak, antzekotasunak eta desberdintasunak.

## 4. Adierazpide geometrikoa

Atal honetan, eraikitako eredua espazio geometriko batetan kokatuko dugu. Kokatze honen erakuntza guztiz orokorra izan liteke beste zenbait eredurentzako.

Bedi $J = \bigcup_{n=1}^{r} X_n$ erantzunen multzoa.

Aipatutako elkartasun maila aplikazio hau da: $k: J \times J \to N$, edozein $(j, j') \in J \times J$ -rentzako

$$k(j,j') = \sum_{w \in \Omega} e^*_j(w)e^*_{j'}(w)$$

Definizio honetatik aise ondorioztatzen dira datozen propietateak:

a. Aplikazio hau simetrikoa da: edozein $(j, j') \in J \times J$ -rentzako $k(j, j') = k(j', j)$

b. Edozein $j \in J$ -rentzako $k(j, j)$ $\Omega$-n izandako bere maiztasuna da.

c. Edozein $j \in J$ eta edozein $X_n$ $n = 1, ..., r$, $k(x_n^+,j) + k(x_n^-,j) = k(j,j)$

d. Edozein $X_n$, $n = 1, ..., r$ -rentzako $k(x_n^+, x_n^-) = 0$

e. Edozein $X_n$, $n = 1, ..., r$ -rentzako $k(x_n^+,x_n^+)+k(x_n^-,x_n^-) = N$

Elkartasun mailak hobeto ulertze arren egokia izaten da bakoitza proporztionalki adieraztea:

edozein $j \in J$ eta edozein $j' \in J$ -entzako:

$f_{j'}^{j} = k(j,j')/k(j,j)$ ; ($f_{j'}^{j} \geq 0$ beraz).

Horrela:

a. edozein $X_n$, $n = 1, \ldots, r$ eta edozein $j \in J$ -entzako:

$$f_{x_{n^+}}^{j} + f_{x_{n^-}}^{j} = 1$$

b. edozein $X_n$, $n = 1, \ldots, r$ -rentzako:

$$f_{x_{n^+}}^{x_{n^+}} = 1; \quad f_{x_{n^-}}^{x_{n^+}} = 0; \quad f_{x_{n^+}}^{x_{n^-}} = 0; \quad f_{x_{n^-}}^{x_{n^-}} = 1.$$

Erantzun baten <u>profila</u> $|R^{2r}$ -ko puntu hau da:

edozein $j \in J$ -rentzako

$$f_J^j = (f_{j'}^{j} | j \cdot J) \in |R^{2r}$$

Erantzun baten <u>masa</u> zenbaki positibo hau da:

edozein $j \in J$ -rentzako

$f_j = k(j,j)/N$.

Bereizgarri hau dute bestalde: edozein $X_n$, $n = 1, \ldots, r$ -rentzako

$$f_{x_{n^+}} + f_{x_{n^-}} = 1; \qquad \sum_{j \in J} f_j = r \quad \text{beraz.}$$

Honela, erantzunen multzoari $|R^{2r}$ -ko puntu-sistema bat egokitu zaio:

$$N(J) = \{(f_J^j, f_j) \mid j \in J\},$$

$f_j^j$ $j \in J$-ren koordenatuak izanik, eta $f_j$ bere masa edo sistema horretan $j \in J$ -ri dagokion garrantzia (bere maiztasunaren arauera).

Puntu-sistema honen grabitate-zentrua (ulerbide mekanikoa) edo batezbesteko profilaren (ulerbide estatistikoa) koordenatuak hauexek dira: edozein $j \in J$ -rentzako:

$$1/r \times \sum_{j' \in J} f_{j'} \times f_j^{j'} = 1/r \times \sum_{j' \in J} \frac{k(j',j')}{N} \times \frac{k(j',j)}{k(j',j')} = 1/r \times \frac{rk(j,j)}{N} = f_j$$

edo laburki esanda: $f_j = (f_j \mid j \in J) \in |R^{2r}$ da $N(J)$-ren batezbesteko profila.

Bestalde, galdera bakoitzaren bi erantzunen profilen batezbestekoa

hauxe bera da ere:

edozein $X_n$, n = 1, ...,r eta edozein j $\in$ J -entzako:

$$f_{x_n^+} \times f_j^{x_n^+} + f_{x_n^-} \times f_j^{x_n^-} = \frac{k(x_n^+,x_n^+)}{N} \times \frac{k(x_n^+,j)}{k(x_n^+,x_n^+)} + \frac{k(x_n^-,x_n^-)}{N} \times \frac{k(x_n^-,j)}{k(x_n^-,x_n^-)}$$

$$= \frac{k(j,j)}{N} = f_j$$

Sistema honen bereizgarri bat beraz, hauxe da: edozein galderaren bi erantzunen profilak eta batezbeteko profila, zuzen batean daude.

Sistema honen barrutik erantzunen arteko erlazioak argitaratu nahi ditugu.

Ohar daitekenez, bi galderaren artean baldintza bikoitzazko erlazioa edo bateraezinezko disjuntziozkoa baldin badago, orduan beren erantzunen profilak binaka hartuta nolabait, berdinak dira. Eta alderantziz, bi erantzunen profilak berdinak baldin badira, beren aurkakoak bai eta ere, eta beraz, dagozkien bi galderen arteko erlazioa baldintza bikoitzazkoa da (baiezkoak edota ezezkoak badira berdinak) ala bateraezinezko disjuntziozkoa (bestela).

Honela esan dezakeguna zera da: bi erantzunen hurbiltasuna espazio honetan, beren baliokidetasunaren legez dagoela burututa. Hurbiltasuna ordea, ez dago oraindik zeharo definituta: bi profilen arteko distantzia definitu behar da erantzunak espazio metriko batetan murgiltzeko.

Halako distantzia bati bereizgarri hau betetzea eskatu lekioke: bi galdera baldintza bikoitzazko erlazioan badaude (edo bateraezinezko disjuntziozkoan) orduan bi galderak batutzeak (espazioaren dimentsioak laburtzen) ez ditu aldatu behar gainontzeko erantzunen arteko distantziak.

Eskakizun honi euklidearra izatearena gehitzen baldin bazaio (ezaguna zaigun errealitate bati lotzeko, eta gainera, erosotasun matematiko bat lortzeko), orduan holako distantzia bat definitzen da:

edozein (j', j'') $\in$ J $\times$ J-rentzako

$$d^2(j',j'') = \sum_{j \in J} (f_j^{j'} - f_j^{j''})/f_j$$

Ikusten denez, batugaiak erantzunen garrantziaz daude neurtuta. Distantzia honi $\chi^2$-ren distantzia deritzaio, bi galderaren kasuan estatistika

klasikoan $\chi^2$-ren testarena gogoratzen duelako.

Erantzunen adierazpide geometrikoa azalduz gero, datorrenean bere analisirako bi metodo matematiko orokor aurkeztuko ditugu: faktore-analisia eta sailkaketa [3].

## 5. Faktore-analisia.

Puntu sistema baten faktore-analisiaren metodoa mende honetan burutu zen psikometriaren alorrean Thurstone-n eskutik (1930. urtearen geroztik), adimen orokorrari buruzko Spearman-en teoriak formalizatuz.

Ikuspuntu asko izan da alor honetan, baina datuen analisiarenatik honela aurkeztu liteke: $N(J)$ puntu-sistemarentzako ardatz nagusia bilatu $\mathbb{R}^{2^r}$ espazioan, hau da, erantzunak edozein ardatzean proiektatu (definitutako distantziaren zentzuan), eta inertzia (ulerbide mekanikoa) edo bariantza (ulerbide estatistikoa) haundiena duena bilatu (karratu minimoen hastapena aplikatu); ondoren, ardatz honekiko perpendikularren artean arazo bera planteatu, eta honelaxe aritu espazioaren dimentsioa agortu arte.

Erantzunek ardatz hauetan hartzen dituzten balioei faktoreak deritzaie, eta nolabait esateko, beren koordenatu berriak dira $\mathbb{R}^{2^r}$ espazioan. Koordenatu cartesiar hauek bereizgarri batzuk dituzte; esaterako, gehienetan faktoreen arteko lehenengo batzuren artean ia informazio guztia (bariantzaren zentzuan) mamitzen da.

$\Omega$ -ko elementuak ere proiektatu daitezke espazio honetan (beraz, ardatz nagusietan), beren profil hauek definituaz:

$$\text{edozein } w \in \Omega \text{ eta edozein } j \in J \text{ -entzako } f_j^w = e^*_j(w).$$

Galdeketaren ereduan faktore-analisia k: $J \times J \to N$ aplikazioan oinarritu beharrean, k': $\Omega \times J \to \{0, 1\}$ aplikazioan oinarritu bagenu, edozein $(w,j) \in \Omega \times J$-rentzako $k'(w,j) = e^*_j(w)$ izanik, J-ko elementuen profilak, masak eta elkarren arteko distantziak arestian definitutakoaren antzeko era batez definituta ( $\Omega$-ko elementuekiko), hala nola $\Omega$ -koenak (J-ko elementuekiko), badirudi erantzunen beste faktore-analisi bat lortuko genukela. Izatez, bi analisiak antzekoak dira eta badute lotura matematikorik. Gainera, $\Omega$-ko elementuen faktoreak berdinak dira bi analisietan.

Frogatu egiten da ere, bigarren faktore-analisian $N(\Omega)$ eta $N(J)$ puntu-sistemen ardatz nagusiak berberak direla, eta beraz, batera adierazi daitezkela erantzunak eta objetuak.

Praktikan emaitza hauek garrantzi haundikoak dira: alde batetik, ardatz bakoitzean zein objetu zein erantzunetik daude hurbil ($\chi^2$-ren distantziaren legez) begiratu liteke eta honela elkartasun maila erlatiboak neurtu; bestetik, nahikoa da emaitzak lortzeko lehenengo analisiari ekitea, honen kalkuluen dimentsioak askoz laburragoak izaten dira eta.

Erantzunen puntu-sistema ardatz nagusi batetan (edota bi ardatz nagusik osatzen duten planu batetan) propiektatzen denez gero, erantzunen arteko antzekotasunak eta, egokiro azaldu daitezke. Adibidez, galderen artean Guttman-en eskala baten egitura balego, lehenengo faktoreak ordenatuko lituzke erantzun guztiak; are gehiago, gainontzeko faktoreak lehenengoaren funtzio polinomikoak izango lirateke.

Faktoreak kalkulatzeko, r-ren balioa oso txikia den kasutan ezik, ordenagailuen erabilpena ezinbestekoa gertatzen da; eskuz egitea zeharo latza eta luzea izango zen.

## 6. Sailkaketa

Sailkaketa metodoen helburua multzo bateko elementuak sail homogeno gutxitan elkartzea da.

Oraingo honetan elementu bakoitza sail bakar batetan dagoen kasutara mugatuko gara, partiketen[4] kasutara alegia.

Metodo hauek bitan banatu litezke: hierarkikoak eta ez hierarkikoak.

Metodo ez hierarkikoek partiketa bat osatzen dute zuzenean, sail kopurua zein den aldez aurretik erabakiz gero. Metodo hierarkikoek berriz, partiketa segida bat osatzen dute, partiketak fintasunaren[5] arauera ordenatuta daudelarik.

Esan daiteke, azken metodo hauek XVII. eta XVIII. mendeetako taxonomilariek burutu zituztela.

Metodo hauen adibide bat, ondoko algoritmo xinple bezain emankorrean oinarritzen da:

1.- Sail bakunen partiketa hartu, alegia, multzoaren elementu bakoitzak sail bat osatzen duenarena.

2.- Bi sail batu erizpide bati jarraituz, eta beraz, partiketa berri bat eraiki (noski, zaharra finagoa da berria baino).

3.- Partiketa berria multzo osoarena baldin bada, gelditu; bestela, 2. pausora itzuli.

Algoritmo hau eta antzekoak aplikatzerakoan, auzia erizpidean dago. Kontutan hartuaz partiketa baten aurrean, alde batetik sailen barruko inertzia edo bariantza kalkulatu daitekela, eta bestetik sailen artekoa (sail bakoitzeko ordezkari bat definituz gero), erizpide bat (Ward (1963)--ena deritzana) hauxe izan liteke: sailen barruko (arteko) bariantza txikiena (haundiena) emango duten bi sailak batu.

$\Omega$ eta J multzoen gain sailkaketa bana eraiki bada, sail bakoitzaren ordezkarien arteko elkartasun mailak azalduaz, hasierakoen (partiketa bakunen) kondentsazio bat lortuko da, eta informazioa mamituta agertuko. Multzo baten eta bestearen sailak ("bereziak" izendatu ditugu idazki honen hasieran) elkarren artean konparatu eta interpretatu litezke. Honela $\Omega$ -ren parte edo sail bateko elementu guztiek era berean erantzun baldin badiete galdera batzuri (erantzunek J-ren sail "berezi" bat osatzen dutelarik), orduan esan liteke $\Omega$ -ren sail horretan konjuntziozko erlazio bat betezen dela.

Baldintzazko erlazioen arazoak leku berezi bat dute sailkaketaren metodoen barruan. $\Omega$ -ko elementuen partiketa bat aldez aurretik ezartzen bada, hain zuzen galdera batek (edo gehiago) definitzen duena, bere sailen eta gainontzeko galderen sailen artean halako erlazioak bilatu daitezke. Estatistikan, arazo hau bereizketa[6] izenaren bidez da ezaguna.

Azkenik, sailkaketa eta faktore-analisiaren arteko lotura dugu. $N(J)$ puntu-sistema aztertzeko distantzia berbera erabili baldin bada bi metodoetarako, biok elkarren osagarriak gertatzen dira, hau da, faktoreak (erlazio nabarmenenak) eta sailak elkarren artean interpretatzeko ezagutzen dira zenbait adierazle. Xehetasun hauetan, hala nola idazki honetan zehar ikutu baizik ez ditugun bestetan sartu nahi duen irakurleak, tresna aberatsak aurkitu litzake ondoko bibliografian.

## OHARRAK

[1] Bietako galderaren bat, eta bakar bat, alderantziz planteatuz gero, baldintza bikoitzazko erlazioa azaltzen da.

[2] Delako Guttman-en eskala bat r galderaren gain (r-1) baldintzazko erlazioa xinple dira, orden egoki batetan horrelako egitura dutelarik:

$$(X_1 \rightarrow X_2) \cap (X_2 \rightarrow X_3) \cap \ldots \cap (X_{r-1} \rightarrow X_r).$$

[3] Factor analysis and classification.

[4] Bitez A eta $\mathbb{P}$ bi multzo. $\mathbb{P}$ A-ren partiketa bat da, baldin eta soilik baldin:

a. edozein $P \in \mathbb{P}$ -rentzako, $P \subset A$ y $P \neq \emptyset$;

b. edozein $P \in \mathbb{P}$ eta edozein $P' \in \mathbb{P}$ -entzako, $P \neq P'$ izanik, $P \cap P' = \emptyset$ bada;

eta

c. $\bigcup_{P \in \mathbb{P}} P = A$;

[5] Bitez $\mathbb{P}$ eta $\mathbb{P}'$ A multzoaren bi partiketa. $\mathbb{P}$, $\mathbb{P}'$ baino finagoa da baldin eta soilik baldin edozein $P \in \mathbb{P}$ -rentzako badago $P' \in \mathbb{P}$, $P \subset P'$ izanik.

[6] discrimination.

## BIBLIOGRAFIA

BENZECRI, J. P. *Histoire et Préhistoire de l'Analyse des Données*, París: Dunod, 1982.

BENZECRI, J. P. et al. *L'Analyse des Données* (Tome I: *La Taxinomie*; Tome II: *L'Analyse des Correspondances*), París: Dunod, 1ère éd. 1973; 4ème éd. 1982.

TUKEY, J. W. "The Future of Data Analysis", *Annals of Mathematical Statistics*, Vol. 33 (1962).

TUKEY, J. W. *Exploratory Data Analysis*, Addison Wesley, 1977.

Facultad de Filosofía y CCEE
Universidad del País Vasco, San Sebastián