Laginketa-erroreen kalkuluari buruzko txostena

Informazioaren Gizarteari buruzko Inkesta
(IGI- Familiak)

AURKIBIDEA

1. Sarrera

Laginketa-errorea definitzean zera esan dezakegu: aztergai den biztanleriaren ezaugarri bat (parametroa) biztanleria horren atal edo lagin batetik ateratako balioaren bidez (estatistikoa) zenbatestean egiten den zehazgabetasuna dela.

Errore hori faktore askoren mende dago, besteak beste, biztanleriaren atal hori hartzeko prozeduraren (laginaren diseinua), hartzen diren unitate kopuruaren (laginaren tamaina), zenbatetsi nahi den ezaugarria nolakoa den, eta abarren mende. Hona hemen laginketa-errorearen adierazpide bat, oso zabalduta dagoena:

$$\text{Error de muestreo} = \sqrt{Var(\hat{\theta})} \qquad (1)$$

Kasu horretan, $\hat{\theta}$ interesaren gaineko estatistikoa da (batezbestekoa, guztirakoa, proportzioa…). Estatistiko horrek balio desberdinak hartuko ditu, ateratako laginaren arabera. Estatistikoak laginketan duen aldakortasunaren mende egongo da laginketa-errorea.

Errore horren adierazpena desberdina izango da laginketarako erabili den teknikaren arabera; izan ere, zenbat eta konplexuagoa izan laginaren diseinua hainbat eta zailagoa izango da kalkulua. Gainera, informazioa biltzean izaten diren gorabeheren, biztanleriaren ezaugarri jakin batzuetarako egokitzapenaren (estratifikazio osteko fasea) eta inkesta burutzen den artean dauden beste faktore batzuen ondorioz, jasokarien eta azken pisuen kalkuluan aldaketak egin behar dira.

Literaturak laginketa-erroreak kalkulatzeko ohiko metodoen aldean zenbait alternatiba iradoki izan ditu. Teknika heuristiko horien bidez laginketa-errorearen zenbatespen ona egin daiteke, laginaren diseinuaren azken pisuak eta ezaugarriak aintzat hartuta [3], [5].

Hemendik aurrera aipatu metodoak sartuko ditugu eta 2000. urtetik aurrera Informazioaren Gizarteari buruzko Inkestan egindako aplikazio zehatza.

2. Taylor-en hedapen-metodoa [3], [5].

Metodo honen bidez laginketa-errorearen zenbatespenak kalkula daitezke guztirakoetarako, batez bestekoetarako eta ehunekoetarako estratifikazioa, klusterra eta probabilitate ezberdina duten laginketetan; hau da, EUSTATek egindako estatistika-eragiketa askotan. Metodoak zenbateslearen hurbilketa linealak ateratzen ditu, eta bere bariantza kalkulatzen du horren laginketa-errorearen zenbatesle bezala erabiliz.

Biztanleriaren batez besteko bezala zenbatetsi den bariantza kalkulatzeko adierazpena ondokoa da:

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} \frac{n_h(1-f_{h)}}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2 \qquad (2)$$

Non:

$$e_{hi.} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij} - \hat{\bar{Y}})}{w...}$$

$$\bar{e}_{h..} = \frac{\sum_{j=1}^{n_h} e_{hi.}}{n_h}$$

eta

$$w... = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}$$

Idazkera:

$h = 1, 2, \ldots, H$ geruza da eta guztira H geruza daude.
$i = 1, 2, \ldots, n_h$, h geruzan dagoen kluster kopurua da, eta guztira n_h kluster daude.
$j = 1, 2, \ldots, m_{hi}$, h geruzaren i klusterraren barruko unitate zenbakia da, eta guztira m_{hi} unitate daude.

lagineko behaketa kopuru osoa da.

$w_{hij,}$ j behaketak h geruzaren i klusterrean duen jasokaria da.
$y_{hij} = (y_{hij}(1), y_{hij}(2), \ldots, y_{hij}(P))$, h geruzako i klusterraren j behaketan Y aldagaiaren gainean ikusi diren balioak dira (zenbaki eta kategoria aldagaiak).

SAS [4] estatistika-multzoko PROC SURVEYMEANS prozedurak laginketa-erroreen zenbatespenerako metodoa aplikatzen du, eta eragiketa honetan laginketa-erroreak kalkulatzeko erabiliko dugu.

3. IGI-Familiak inkestako erroreen kalkulua.

3.1 Lagin-diseinua [1]

Informazioaren Gizarteari buruz familiei egindako Inkesta (IGI-Familiak) izeneko inkesta laginketa bidezkoa da eta Euskal AEko 6 urteko eta gehiagoko biztanleak hartzen ditu aztergai. Inkesta honen laginketarako oinarri bezala, hiruhileko berean Biztanleria jardueraren arabera sailkatzeko inkestarako (BJA) hautatu ziren etxebizitzen laginketa panela hartu da. 2004. urtetik aurrera, laginean 5.088 etxebizitza zeuden eta Etxebizitzen Direktoriotik ateratzen da ausaz, lurralde historikoaren araberako modu estratifikatuan. Geruza bakoitzaren barruan etxebizitzaren laginak sistematikoki ateratzen dira (probabilitate berarekin) [2].

Etxebizitza bakoitzean lehenengo pertsona ausaz hautatzen da, Kish-en taula baten bidez eta, gainera, okupatuak edo ikasleak daudenean horietariko bana ere hautatuko da prozedura bera erabiliz. 2003. urtetik aurrera, lagina osatzeko, 6 eta 14 urte bitarteko adingabeko guztiak hartzen dira, gutxi gorabehera 7.500 pertsonako lagina izan arte.

Urtero ustiatzen da inkesta (aztertzen den urteko 1. hiruhilekoa) eta emaitzak norbanakoei eta familiei buruzkoak dira, baina tratamendu berezia ematen zaie Interneteko erabiltzaileei.

Deskribatu den diseinu hori ezin hobeto egokitzen da aurreko idatz-zatian azaldu den metodo heuristikoaren zehaztapenetara. SASaren prozedurak eskatzen dituen parametroak besterik ez dira adierazi beharko, bariantza behar bezala zenbatetsi ahal izateko.

3.2 Kalkulurako prozedura.

Hona hemen erroreen kalkulurako aplikatu den SASaren prozeduraren oinarrizko sintaxia [4]:

PROC SURVEYMEANS < *fitxategiaren_izena* > < *irteerarako aukerak* >;
BY *aldagaiak* ; /*erroreen kalkulua, azpipopulazio independenteen arabera*/
CLASS *aldagaiak*; /*erroreen kalkulua, aldagai kualitatiboetarako*/
CLUSTER *aldagaiak* ; */laginketan dagoen klusterra adierazten duen aldagaia, konglomeratuen arabera*/
DOMAIN *aldagaiak* ; /*erroreak kalkulatu nahi diren eremua/gurutzaketa mugatzen duten aldagaiak*/
RATIO *aldagaia/aldagaia* ; /*laginketa-errorea kalkulatu nahi den ratio-aldagaiak*/
STRATA *aldagaiak* < / *option* > ; /*estratifikatutako laginketan geruza adierazten duen aldagaia*/
VAR *aldagaiak* ; /* laginketa-erroreak kalkulatu nahi diren aldagai kuantitatibo eta kualitatiboak*/
WEIGHT *aldagaia* ; /* aurretik kalkulatutako pisu-aldagaia (aukerazkoa)*/

IGI – Familiak inkestarako sintaxi honen parametro orokorrak ondokoak izango dira:

CLUSTER = etxebizitzaren identifikatzailea.
STRATA = lurralde historikoa.
WEIGHT = Pertsonen urteko jasokaria /familien urteko jasokaria.
VAR = Informazioaren teknologien ekipamendu eta erabileraren aldagaiak.
DOMAIN = gurutzaketak, aldagai sozio-demografikoen eta ekonomikoen arabera.

3.3 IGI-Familiak inkestako erroreak kalkulatzeko estatistikoak eta eremuak.

Hurrengo gurutzaketa eta estatistikoetarako laginketa-erroreak zenbatetsiko dira:

Familiak

- Euskal AEko familiak IKT ekipamenduen arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Euskal AEko familiak IKT ekipamenduen arabera, familia motako (%) 2011. Laginketa-erroreak.

- Euskal AEko familiak etxeko telebista ekipamenduen arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.

- Euskal AEko familiak etxeko telebista ekipamenduen arabera, familia motako (%) 2011. Laginketa-erroreak.

Biztanleria

- Euskal AEko 15 urte eta gehiagoko biztanleria etxeko IKT eta telebista ekipamenduen arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Etxean ordenagailua duen Euskal AEko 15 urte eta gehiagoko biztanleria sexuaren eta adinaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Etxean ordenagailua duen Euskal AEko 15 urte eta gehiagoko biztanleria heziketa mailaren eta jarduerarekin duen harremanaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Etxean internet duen Euskal AEko 15 urte eta gehiagoko biztanleria sexuaren eta adinaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Etxean internet duen Euskal AEko 15 urte eta gehiagoko biztanleria heziketa mailaren eta jarduerarekin duen harremanaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Euskal AEko 15 urte eta gehiagoko biztanleria Lurralde historikoko eta ezaugarri demografikoekiko, etxeko IKT ekipamenduen arabera (%) 2011. Laginketa-erroreak.
- Euskal AEko 15 urte eta gehiagoko biztanleria internet eskura izateko aukeren eta Lurralde Historikoaren arabera (%) 2011. Laginketa-erroreak.

Interneten erabilera

- Internet erabiltzen duen Euskal AEko 15 urte eta gehiagoko biztanleria sexuaren eta adinaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Internet erabiltzen duen Euskal AEko 15 urte eta gehiagoko biztanleria heziketa mailaren eta jarduerarekin duen harremanaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Internet erabiltzen duen Euskal AEko 15 urte eta gehiagoko biztanleria erabiltzen dituen zerbitzuen eta azkeneko konexioaren batez besteko iraupenaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Internet erabiltzen duen Euskal AEko 15 urte eta gehiagoko biztanleria sartzen den tokiaren eta erabiltzen duen hizkuntzaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.
- Internetez hondasunak erosi dituen Euskal AEko 15 urte eta gehiagoko biztanleria, ordaintzeko moduaren eta ordaintzeko segurtasunari buruzko iritziaren arabera, Lurralde Historikoko (%) 2011. Laginketa-erroreak.

- Euskal AEko 15 urte eta gehiagoko biztanleria konexioaren amaieraren, sartzeko maiztasunaren, konexioaren iraupenaren eta Lurralde Historikoaren arabera (%) 2011. Laginketa-erroreak.
- Euskal AEko 15 urte eta gehiagoko biztanleria erabilitako zerbitzuen, ikusitako web moten eta Lurralde Historikoaren arabera (%) 2011. Laginketa-erroreak.

Hurrengo tauletan arestian aipatutakoa laburtzen da, estatistikoaren eta gurutzaketa-aldagaiaren arabera:

Estatistikoa	IKT ekipamenduak	Lurralde historikoa	Sexua	Adina	Hezketa maila	Jarduerarekiko lotura	Familia mota	Internet ekiko erlazioa	Merkataritza elektronikoa
15 urteko eta gehiagoko biztanleriaren ehunekoa	X	X	X	X	X	X	X		
15 urteko eta gehiagoko biztanleak, guztira (milakotan)	X	X	X	X	X	X	X		
Internet erabiltzen duten 15 urteko eta gehiagoko biztanleen ehunekoa		X	X	X	X	X		X	X
Internet erabiltzen duten 15 urteko eta gehiagoko biztanleak, guztira (milakotan)		X	X	X	X	X	X	X	X
Familien ehunekoa (%)	X	X					X		
Familiak guztira (milakotan)	X	X					X		

3.4 Emaitzak eta interpretazioa.

Laginketa-errorea zenbatesteaz gainera (2), SASak baliagarriak diren eta errorea interpretatzen laguntzen duten beste neurri batzuk ere ematen ditu. Hona hemen, besteak beste, interesgarrienak:

- **Aldakuntza-koefizientea.** Zehaztapenak multzo edo populazio desberdinen artean alderatzen lagatzen duen errorearen neurri erlatiboa da. Dimentsiorik ez duen magnitudea da, laginketa-errorearen neurri gisa oso erabilia; hauxe du adierazpidea:

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{\hat{\theta}} \qquad (3)$$

- **Konfiantza-tartea %95era.** Konfiantza-tarte hau estatistikoaren laginketaren banaketan oinarritzen da (proportzioa, batez bestekoa, tasa,…). Limitearen Teorema Nagusiaren bidez, gehien-gehienetan lege Normal[1] bat onar dezakegu estatistikorik ohikoenetarako; beraz, tarte honen eraikuntza honako adierazpide honen ondorio izango da:

$$\left[\hat{\theta} - 1{,}96\sqrt{Var(\hat{\theta})},\quad \hat{\theta} + 1{,}96\sqrt{Var(\hat{\theta})}\right] \qquad (4)$$

1,96 balioa % 95eko probabilitatea daukan batez bestekoa 0 eta desbiderapen tipikoa 1 dituen banaketa Normal baten pertzentila da. Horrela baiezta daiteke $\hat{\theta}$ estatistikorako kalkulatutako tarteak biztanleriaren parametroaren benetako balorea daukala kasuen %95ean (balizko laginak).

SASak emandako informazioarekin, estatistikoaren zenbatespena, %95eko konfiantza-tartearen beheko eta goiko mugak eta aldakuntza-koefizientea (ehunekoa) jasoko duten erroreen behin betiko taulak egingo dira.

Jarraian erroreak zabaltzeko taularen eredu bat ageri da:

Euskal AEko familiak etxeko IKT ekipamenduen arabera, Lurralde Historikoko (%) 2011. Laginketako akatsak.

	Guztira (milakotan)	Ordenagailua	Internet	Eskuko telef.	Helb. elek.
Euskal AE					
Zenbatespena	846,3	62,4	57,0	90,1	55,1
% 95etik beherako m.	842,8	61,0	55,5	89,2	53,7
% 95etik gorako m.	849,8	63,8	58,4	90,9	56,6
AK (%)	0,2	1,1	1,3	0,5	1,3

Iturria: EUSTAT. Informazioaren Gizarteari buruzko Inkesta-IGIF.

Informazio hau interpretatzeko beste modu bat konfiantzaren % 95ari dagokion errorea kalkulatzea da; 1,96 pertzentila aldakuntza-koefizienteaz bideratuz ateratzen da hori**.** Errore erlatibo honek zenbatespenaren balioaren ehuneko puntutan hitz egiteko aukera ematen digu.

Aurreko taularako, Euskal AEn ordenagailua duten familien ehunekoaren % 95erako errore erlatiboa % 2,2 da (1,96*1,1). Edo, bestela esanda, % 95eko konfiantza mailan Euskal AEn ordenagailua duten familien ehunekoaren benetako balioa emandako zenbatespenaren ±%2,2ko tartean dabil. Hau da,

(62,4 ± 0,02156*62,4) = % 61,0 eta % 63,8 artean

Garrantzitsua da % 95erako errore erlatiboaren ehuneko jakin bat gainditzen duten zenbatespenak zein diren aipatzea. Zentzuzko muga akats erlatiboaren % 20 gainditzen duten zenbatespenetan legoke (G.K. > % 10 gutxi gorabehera), errore hori % 30etik gorakoa zen laukitxoak bereziki azpimarratuz (G.K. > % 15 gutxi gorabehera).

Bibliografia

[1] EUSTAT (2005), "*Informazioaren Gizarteari buruzko Inkesta. IGI – Familiak*. Fitxa metodologikoa http://www.eustat.es/document/esi_c.html

[2] EUSTAT (2005), "*Biztanleria jardueraren arabera sailkatzeko inkesta*. *Ohar metodologikoa. 2005"* http://www.eustat.es/document/datos/notamet_nuevaPRA_c.pdf

[3] Fuller, W. A. (1975), *"Regression Analysis for Sample Survey,"* Sankhy , 37, Series C, Pt. 3, 117 - 132.

[4] Sas Institute Inc. (2004), *"SAS/STAT® 9.1 Guia de Usuario".* Copyright © 2004, Cary, NC, USA. ISBN 1-59047-243-8

[5] Woodruff, R. S. (1971), "*A Simple Method for Approximating the Variance of a Complicated Estimate"* Journal of the American Statistical Association, 66, 411 -414.