

Laginketa-erroreen kalkuluari buruzko txostena

Biztanleria jardueraren arabera sailkatzeko inkesta
(BJA)



AURKIBIDEA

1. Sarrera.....	3
2. Taylor-en hedapen-metodoa.....	3
3. Erroreen kalkulua BJA.....	4
3.1 Laginaren diseinua	4
3.2 Kalkulurako prozedura.....	5
3.3 BJA inkestako erroreak kalkulatzeko estatistikoak eta eremuak.	5
3.4 Emaitzak eta interpretazioa	7
Bibliografia.....	9

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 da.

Errore hori faktore askoren mende dago, besteak beste, biztanleria 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:

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

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

Errore horren adierazpena desberdina izango da laginketarako erabili den teknikaren arabera; izan ere, zenbat eta laginaren diseinua konplexuagoa izan hainbat eta zailagoa izango da kalkulua. Gainera, informazioa biltzean izaten diren gorabeheren, populazioaren 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 lagin-errorearen zenbatespen ona egin daiteke, laginaren diseinuaren azken pisuak eta ezaugarriak aintzat hartuta [3], [5].

Ondoren, metodo horiek emateaz gainera, Biztanleria jardueraren arabera sailkatzeko inkestan (aurrerantzean BJA) 2005. urtetik aurrera izan duten aplikazio zehatza ere ematen da.

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

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

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

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

Kasu horretan:

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

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

eta

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

Idazkera:

$h = 1, 2, \dots, H$ geruza da eta guztira H geruza daude.

$i = 1, 2, \dots, n_h$, h geruzan dagoen kluster kopurua da, eta guztira n_h kluster daude.

$j = 1, 2, \dots, m_{hi}$, h geruzaren eta i klusterraren barruko unitate kopurua da, eta guztira m_{hi} unitate daude.

$n = \sum_{h=1}^H \sum_{i=1}^{n_h} m_{hi}$ lagineko behaketa kopuru osoa da.

w_{hij} , j behaketak h geruzaren i klusterrean duen jasokaria da.

$y_{hij} = (y_{hij}(1), y_{hij}(2), \dots, y_{hij}(P))$, 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 estatistika-eragiketa honetan laginketa-erroreak kalkulatzeko erabiliko dugu.

3. Erroreen kalkulua BJAren

3.1 Lagin-diseinua [1]

2005. urtean, BJAren laginketa aldatu egin zen; ondorioz, aurreko aldietan baino unitate gehiago egoteaz gainera diseinu aldetik konplexutasuna txikiagoa da orain [2]. Egungo laginketak probabilitate-lagin jarraitua hartzen du oinarri gisa, hau da, etengabe berritzen doan etxebizitzaren panela. Laginean gutxi gorabehera hiruhilekoan 5.000 etxebizitza sartzen dira (guztira 13.500 norbanako inguru), eta zortzigarren bat errotatuz doa hiruhileko batetik bestera: hartara, etxebizitza bakoitza laginean bi urtez egoten da.

Etxebizitzaren lagina Etxebizitzaren Direktoriotik ateratzen da ausaz, lurralde historikoaren arabera modu estratifikatuan. Geruza bakoitzaren barruan etxebizitzaren laginak sistematikoki ateratzen dira (probabilitate berarekin). Informatzaile baten bitartez etxebizitzako pertsona guztiei buruzko informazioa jasotzen da eta, horrela, diseinuari dagokionez etxebizitzak konglomeratu txiki bezala jardungo dute. Etxebizitzak dira hautatutako unitateak eta emaitzak norbanakoei buruzkoak dira. Diseinu hori ezin hobeto egokitzen da aurreko idatz-zatian deskribatu den metodo heuristikoen zehaztapenetara. Soilik adierazi beharko dira, beraz, SASaren prozedurak eskatzen dituen parametroak bariantza behar bezala zenbatetsi ahal izateko.

3.2 Kalkulurako prozedura.

Hona hemen erroreen kalkulurako aplikatu den SASaren prozeduraren oinarritzko 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 ; /*laginketaren errorea kalkulatu nahi den ratio-aldagaiak*/  
  STRATA aldagaiak < / option > ; /*estratifikatutako laginketan geruza adierazten duen  
aldagaia*/  
  VAR aldagaiak ; /* laginketaren erroreak kalkulatu nahi diren aldagai kuantitatibo eta  
kualitatiboak*/  
  WEIGHT aldagaia ; /* aurretik kalkulaturako pisu-aldagaia (aukerazkoa)*/
```

BJA berrirako soilik, sintaxi honen parametro orokorrak ondokoak izango dira:

CLUSTER = etxebizitzaren identifikatzailea.

STRATA = lurralde historikoa.

WEIGHT = pertsonen hiruhekoko jasokaria /familien hiruhekoko jasokaria
Pertsonen urteko jasokaria /familien urteko jasokaria.

RATIO = langabezia-tasak, jarduera eta lanbidea.

VAR = langabezian eta lanean dauden biztanleak, biztanle aktiboak... guztira.

DOMAIN = gurutzaketak, aldagai sozio-demografikoen eta ekonomikoen arabera. (ikusi 3.3 idatz-zatia)

3.3 BJA inkestako erroreak kalkulatzeko estatistikoak eta eremuak.

Hurrengo gurutzaketa eta estatistikoetarako laginketa-erroreak zenbatetsiko dira:

Hiruhilekoak

- 16 urteko eta hortik gorako biztanleen jarduera tasa eta langabezia-tasa, lurralde historikoaren arabera (%).
- 16 urteko eta hortik gorako biztanleen jarduera tasa, sexuaren eta adinaren arabera (%).
- 16 urteko eta hortik gorako biztanleen langabezia-tasa, sexuaren eta adinaren arabera (%).

BIZTANLERIA JARDUERAREN ARABERA SAILKATZEKO INKESTA (BJA)

- 16-64 urte bitarteko biztanleen okupazio-tasa, sexuaren, adinaren eta lurralde historikoaren arabera (%).
- 16 urteko eta hortik gorako biztanleak, lurralde historikoaren eta sexuaren arabera (milakoetan).
- 16 urteko eta hortik gorako biztanle aktiboak, lurralde historikoaren eta sexuaren arabera (milakoetan).
- 16 urteko eta hortik gorako biztanle okupatuak, lurralde historikoaren eta sexuaren arabera (milakoetan).
- 16 urteko eta hortik gorako biztanle okupatuak, lurralde historikoaren eta sektore ekonomikoaren arabera (milakoetan).
- 16 urteko eta hortik gorako langabetuak, lurralde historikoaren eta sexuaren arabera (milakoetan).

Urtekoak

- 16 urteko eta hortik gorako biztanleak, jarduerarekin duten loturaren, lurralde historikoaren eta sexuaren arabera (milakoetan).
- 16 urteko eta hortik gorako biztanle okupatuak, lanbide egoeraren eta kontratu motaren arabera (milakoetan).
- Familiak, euren kideen jarduerarekin duten loturaren eta lurralde historikoaren arabera (milakoetan).

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

Hiruhilekoko erroreak

Estatistikoa/gurutzaketa aldagaia	Lurralde historikoa	Sexua	Adina (3 talde)	Sektore ekonomikoa
Jarduera tasa	X	X	X	
Langabezia-tasa	X	X	X	
Okupatuaren tasa	X	X	X	
Biztanleria aktiboa	X	X		
Biztanleria okupatua	X	X		X
Biztanleria langabea	X	X		
16 urteko eta hortik gorako biztanleria	X	X		

Oharra: Gerizadun gurutzaketak aldi berean gertatzen dira.

Urteko erroreak

Estatistikoa/gurutzaketa aldagaia	Lurralde historikoa	Sexua	Jarduerarekin lotura	Lanbide egoera	Kontratu mota
16 urteko eta hortik gorako biztanleria	X	X	X		
Biztanleria okupatua				X	X
Familiak	X		X		

Oharra: Gerizadun gurutzaketak aldi berean gertatzen dira.

3.4 Emaizak eta interpretazioa.

Laginketako 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 erroreaken neurri erlatiboa da. Dimentsiorik ez duen magnitudea da, laginketa-erroreaken neurri gisa oso erabilia; hauxe du adierazpidea:

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

- **Konfiantza-tartea %95era.** Konfiantza-tarte hau estatistikoaren laginketaren banaketan oinarritzen da (proportzioa, batez bestekoa, tasa,...). Limitearen Teorema Nagusiaren bidez, gehien-gehienetan lege Normal¹ bat onar dezakegu estatistikorik ohikoenetarako; beraz, tarte honen eraikuntza honako adierazpide honen ondorio izango da:

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

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

SASak emandako informazioarekin, estatistikoaren zenbatespena, %95eko konfiantza-tartearen beheko eta goiko mugak eta portzentajeazko aldakuntza-koefizientea jasoko duten erroreen behin betiko taulak egingo dira. Jarraian erroreen zabaltzeko taularen eredu bat ageri da:

T.1 Coeficientes de variación e intervalos de confianza para la tasa de actividad y paro de la población de 16 y más años según el territorio histórico (%). IV-2004

Fuente: EUSTAT. Encuesta de Población en Relación con la Actividad.

	C.A. de Euskadi		Araba / Alava		Bizkaia		Gipuzkoa	
	Tasa de actividad	Tasa de paro	Tasa de actividad	Tasa de paro	Tasa de actividad	Tasa de paro	Tasa de actividad	Tasa de paro
Estimación	55,0	7,0	57,7	5,1	53,3	7,7	56,5	6,7
L. Inferior 95%	53,8	6,2	55,1	3,7	51,6	6,4	54,7	5,5
L. Superior 95%	56,2	7,8	60,2	6,4	55,1	9,0	58,4	8,0
CV(%)	1,1	5,9	2,3	13,5	1,7	8,4	1,7	9,3

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

Aurreko taularako, Euskal Autonomia Erkidegoko Jarduera Tasaren %95erako errore erlatiboa %2,1ekoa da ($1,96 \cdot 1,1$). Edo, bestela esanda, %95eko konfiantza mailan Euskal Autonomia Erkidegoko Jarduera tasaren benetako balioa emandako zenbatespenaren $\pm 2,1$ eko tartean dabil:

$$(55,0 \pm 0,021 \cdot 55,0) = (53,8, 56,2)$$

Garrantzitsua da %95erako errore erlatiboaren portzentaje jakin bat gainditzen duten zenbatespenak zein diren aipatzea. Zentzuzko muga errore erlatiboaren %20 gainditzen duten zenbatespenetan

legoke ($G.K. > \%10$ gutxi gorabehera), errore hori $\%30$ etik gorakoa zen laukitxoak bereziki azpimarratuz ($G.K. > \%15$ gutxi gorabehera).

Bibliografia

- [1] EUSTAT (2005), "Biztanleria jardueraren arabera sailkatze inkesta. Fitxa metodologikoa." http://www.eustat.es/document/poblact_c.html
- [2] EUSTAT (2005), "Biztanleria jardueraren arabera sailkatze inkesta. Ohar metodologikoa.2005." http://www.eustat.es/document/datos/notamet_nuevaPRA_c.pdf
- [3] Fuller, W. A. (1975), "Regression Analysis for Sample Survey," Sankhyā, 37, Series C, Pt. 3, 117 - 132.
- [4] Sas Institute Inc. (2004), "SAS/STAT® 9. "User's Guide". 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.