ESTATISTIKARAKO SARRERA

Koerlazioa, erregresioa eta datu anizkoitzen analisia

Plangintza berriei egokitua

Karmele Fernandez Agirre

udako euskal unibertsitatea

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea

ESTATISTIKARAKO SARRERA

Koerlazioa, erregresioa eta datu anizkoitzen analisia

Plangintza berriei egokitua

Karmele Fernandez Agirre

I.S.B.N.: 84-86967-70-8
Lege-gordailua: BI-19-96

Inprimategia: BOAN S.A. Padre Larramendi 2. BILBO

Banatzaileak: UEU. General Concha 25, 6. BILBO
Zabaltzen: Igarabide, 88. DONOSTIA

AURKIBIDEA

HITZAURREA ..... *XIII*

I. EZAUGARRI ESTATISTIKO BAKUNAK, BANAKETAK, TAULAK, ADIERAZPIDE GRAFIKOAK

I.1. ESTATISTIKAREN HISTORIA LABURRA. ESTATISTIKA ETA PROBABILITATEA *3*
I.2. OROKORTASUNAK *5*
I.2.1. Populazioa eta lagina *5*
I.2.2. Unitate estatistiko edo indibiduoak *6*
I.3. EZAUGARRI ESTATISTIKOAK *6*
I.3.1. Ezaugarri estatistiko bakunak *6*
I.3.1.1. Ezaugarri kuantitatiboak *7*
I.3.1.2. Ezaugarri kualitatiboak *7*
I.4. MAIZTASUN-BANAKETA BAKUNAK *8*
I.4.1. Maiztasun-banaketak (absolutuak, erlatiboak eta metatuak) *8*
I.4.2. Taulak *9*
I.5. MAIZTASUN-BANAKETA BAKUNEN ADIERAZPIDE GRAFIKOAK *13*
I.5.1. Barra-diagramak *13*
I.5.2. Histogramak *15*
I.5.3. Maiztasun-poligonoak *17*
I.5.4. Diagrama linealak *18*
I.5.5. Beste adierazpide grafikoak *18*

II. EZAUGARRI BAKUNEN BALIO TIPIKOAK

II.1. MOMENTUAK *23*
II.1.1. Momentu arruntak edo jatorriarekikoak *23*
II.1.2. Momentu zentratuak edo batezbestekoarekikoak *24*

II.2. MAIZTASUN-BANAKETA BAKUNEN BALIO TIPIKO EDO ESTATISTIKOAK 26
II.2.1. Zentru-joeraren edo posizioaren balio tipikoak 26
II.2.1.1. Batezbesteko aritmetikoa 27
II.2.1.2. Batezbesteko geometrikoa 30
II.2.1.3. Batezbesteko harmonikoa 31
II.2.1.4. Moda 31
II.2.1.5. Mediana eta koantilak 34
II.2.2.Sakabanatzearen balio tipikoak 36
II.2.2.1. Bariantza 37
II.2.2.2. Batezbesteko desbidazioa 38
II.2.2.3. Ibiltartea 39
II.2.3. Itxuraren balio tipikoak 40
II.2.3.1. Asimetri koefizientea 40
II.2.3.2. Kurtosi koefizientea edo zapaltasun/zorroztasun-koefizientea 41
II.3. ALDAGAIEN TRANSFORMAZIO LINEALAK 42
II.3.1. Aldagai zentratua 43
II.3.2. Aldagai tipifikatua..... 44
II.4. KONTZENTRAZIO-NEURKETAK 44
II.4.1. LORENZ-en kurba 45
II.4.2. GINI-ren indizea 47

III. EZAUGARRI ESTATISTIKO BIKOITZAK, BANAKETAK, TAULAK, ADIERAZPIDE GRAFIKOAK

III.1. EZAUGARRI ESTATISTIKO BIKOITZAK 51
III.2. MAIZTASUN-BANAKETA BIKOITZAK 51
III.2.1. Maiztasun-banaketa bikoitzak eta bazter-maiztasunak 51
III.2.2. Maiztasun-banaketa bikoitzak: Taulak 53
III.3. MAIZTASUN-BANAKETA BIKOITZEN ADIERAZPIDE GRAFIKOAK 55
III.3.1. Sakabanatze-diagrama edo puntu-hodeia 55
III.4. MAIZTASUN-BANAKETA BALDINTZATUAK 57

III.5. TAULA BATEN DEPENDENTZIA EDO INDEPENDENTZIA 60
III.6. KONTINGENTZI TAULA BATEN ERRENKADA ETA ZUTABEEN BATEZBESTEKO SOSLAIAK 62

IV. EZAUGARRI BIKOITZEN BALIO TIPIKOAK

IV.1. MOMENTUAK 65
IV.1.1. Momentu arruntak edo jatorriarekikoak 65
IV.1.2. Momentu zentratuak edo batezbestekoarekikoak 66
IV.2. MAIZTASUN-BANAKETA BIKOITZEN BALIO TIPIKO EDO ESTATISTIKOAK 67
IV.2.1. Kobariantza: S_{xy} 67
IV.2.2. Koerlazio-koefizientea: r_{xy} 68
IV.3. ALDAGAIEN TRANSFORMAZIO LINEALAK 70
IV.4. KOERLAZIO GABEKO ALDAGAIEN BI PROPIETATE..... 73
IV.5. SAKABANATZE- ETA KOERLAZIO-MATRIZEAK 74
IV.6. OROKORPENA 75

V. KOERLAZIOA ETA ERREGRESIOA

V.1. SARRERA 79
V.2. ERREGRESIOA R^2-n 79
V.2.1. Batezbestekoaren erregresioa 80
V.2.2. Karratu txikienen erregresioa 80
V.3. KARRATU TXIKIEN ERREGRESIO LINEALA R^2-n 81
V.3.1. Karratu txikienen erregresio zuzena 81
V.3.2. Karratu txikienen erregresio linealaren propietateak 84
V.3.3. Hondar bariantza eta mugatze-koefizientea 85
V.3.4. Erregresio-koefizientearen eta koerlazio-koefizientearen zeinuaren azterketa. Doikuntzaren egokitasuna 86
V.4. ERREGRESIO LINEALA R^n-n 89
V.4.1. Sarrera 89
V.4.2. Erregresio hiperplanoa 90

V.4.3. Propietateak 91
V.4.4. Erregresio hiperplanoaren koefizienteak lortzeko metodoa 92
V.4.5. β erregresio partzial estandardizatuen koefizienteak 94
V.4.6. Hondar bariantza. Mugatze-koefizientea eta koerlazio-koefiziente anizkoitza 95
V.4.7. Edozein aldagai azalduaren erregresioaren orokorpena 97
V.4.8. Erregresio-koefiziente desberdinen zeinuaren azterketa. Doikuntzaren egokitasuna 97
V.5. KOERLAZIO PARTZIALA 98
V.5.1. Koerlazio partziala R^3-n 99
V.5.2. Koerlazio partziala R^n-n 101
V.5.3. Koerlazio partzial eta erregresio-koefizienteen arteko erlazioa 103

V. A. Eranskina: ALDAGAI ESTATISTIKO "n"-KOITZEN MATRIZE-ESTATISTIKOAK

V.A.1. DATU-MATRIZEAK 107
V.A.2. BATEZBESTEKO-BEKTOREA: PROPIETATEAK..... 108
V.A.3. KOBARIANTZA MATRIZEA 109
V.A.3.1. Kobariantza matrizearen propietateak 110
V.A.4. KOERLAZIO-MATRIZEA 112
V.A.4.1. Koerlazio-matrizearen propietateak 114
V.A.5. DERIBAZIO BEKTORIALA 114

VI. ZENBAKI-INDIZEAK

VI.1. INDIZE SINPLEAK 117
VI.2. PONDERAZIO GABEKO INDIZE KONPLEXUAK 118
VI.2.1. Batezbesteko aritmetiko sinplearen metodoa 118
VI.2.2. Batezbesteko agregatu sinplearen metodoa 118

VI.3. INDIZE KONPLEXU PONDERATUAK 124
VI.3.1. Balioen, prezioen eta kopuruen indizeak 124
VI.3.1.1. LASPEYRES-en indizeak 127
VI.3.1.2. PAASCHE-ren indizeak 127
VI.3.1.3. FISHER-en indizeak 129
VI.3.1.4. Propietate eta erlazio batzuk: batera-garritasuna eta alderantzizkotasuna..... 131
VI.3.1.5. Kalkulua 132
VI.4. INDIZE KONPLEXUEN ERAIKETAN SORTZEN DEN ZENBAIT ERAGOZPEN 134
VI.4.1. Aldagaien hautapena 134
VI.4.2. Somatutako leku eta denboraren hautapena 134
VI.4.3. Talde eta azpitaldeen hautapena 134
VI.4.4. Oinarri-denboraren hautapena 134
VI.4.5. Formula eta ponderazioen hautapena 135
VI.4.6. Indizearen esangura eta zabaldura 135
VI.5. ERAGOZPEN BEREZI BATZUK 136
VI.5.1. Oinarri-aldaketa indize sinpleetan 136
VI.5.2. Berriztapen eta loturak indize konplexuetan 136
VI.6. ZENBAKI-INDIZEEN APLIKAZIOAK 138
VI.6.1. Kontsumo-prezioen indizeak 138
VI.6.2. Moneta-unitate arruntetan dauden magnitudeen deflazioa 139

VII. DESKRIBAPEN ESTATISTIKOAREN ADIERAZPIDE GEOMETRIKOAK

VII.1. OROKORTASUNAK 143
VII.2. X DATU-MATRIZEAREN BI ADIERAZPIDE GEOMETRIKO 143
VII.3. BI ALDAGAI ETA BI OHARPEN 144
VII.4. BI ALDAGAI ETA HIRU OHARPEN..... 151
VII.5. ERREGRESIO BAKUNA ALDAGAI ESTATISTIKOEN BALIO-MULTZOEN ESPAZIOAN 157

VIII. DOIKUNTZA ORTOGONALA ETA KOBARIANTZA MATRIZEAREN AUTODIREKZIOAK

VIII.1. SARRERA 163
VIII.2. ANALISIA $\mathbb{R}^2$ ESPAZIOAN. DOIKUNTZA ORTOGONALAREN ZUZENA 163
VIII.3. DOIKUNTZA ORTOGONALAREN ZUZENAREN LORPENA 167
VIII.4. HODEI DUALAREN AURKEZPEN GRAFIKOA 170
VIII.5. TRANTSIZIO ERLAZIOAK 172
VIII.6. BATERAKO AURKEZPEN GRAFIKOA ETA INTERPRETAZIOA 173
VIII.7. DOIKUNTZA ORTOGONALAREN ZUZENA ALDAGAI NORMATUENTZAKO 174

IX. DATU ANIZKOITZEN ANALISIA

IX.1. SARRERA 181
IX.2. ANALISI OROKORRA 182
IX.2.1. $\mathbb{R}^n$ espazioaren $\mathbb{R}^q$ azpiespazio baten bidez egindako doikuntza 183
IX.2.2. $\mathbb{R}^m$ espazioaren $\mathbb{R}^q$ azpiespazio baten bidez egindako doikuntza 186
IX.2.3. $\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioen arteko erlazioa 187
IX.2.4 X datu-taularen berreraketa 189
IX.3. OSAGAI NAGUSIZKO ANALISIA 191
XII.3.1. $\mathbb{R}^n$ espazioan egindako analisia 192
XII.3.2. $\mathbb{R}^m$ espazioan egindako analisia 193
XII.3.3. Baterako aurkezpen grafikoak 197
XII.3.4. Adibidea 199
***IX**.4. KORRESPONDENTZI ANALISI FAKTORIALA 207*
IX.4.1. Hodeiak, masak eta distantziak 210
IX.4.2. $\mathbb{R}^n$ espazioan egindako analisia 213
IX.4.3. $\mathbb{R}^m$ espazioan egindako analisia 214
IX.4.4. $\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioen arteko erlazioa 215

IX.4.5. Maiztasun taularen berreraketa 217
IX.4.6. Interpretaziorako laguntzak 218
IX.4.7. Korrespondentzia Anitzeko Analisi Faktoriala 220
IX.5. ELEMENTU GEHIGARRIAK 223
IX.6. SAILKAPEN-METODOAK 224
IX.6.1. Goranzko Sailkapen Hierarkikoa 224
IX.6.1.1. Bariantzaren metodoa 225
IX.6.2. Adibidea 227

HITZAURREA

Eskuartean daukazun liburua, Ekonomia eta Enpresa-Zientzien Fakultateko "Estatistikarako Sarrera" asignaturaren testuliburu gisa, sortuz eta moldatuz joan garen liburu batzuren ondotik dator.

Duela 13 urte, 1983. urtean alegia, *Estatistikaren Hastapenak* izenaz lehen apunte-liburua argitaratu genuen, eta ondoren, *Estatistika Deskribatzailea. Koerlazioa, Erregresioa eta Datu Anizkoitzen Analisia* izenburuaz 1989. eta 1993. urteetan asignaturaren egitarauei egokituz eta gai batzuk berrituz bi argitalpen atera genituen.

Orain, bi urte t'erdi igaro ondoren *Estatistikarako Sarrera. Koerlazioa, Erregresioa eta Datu Anizkoitzen Analisia* kaleratzen dugu. Bi urte t'erdi epe laburra da, baina Sarrikoko ikastetxean aldaketa asko gertatu dira. Hauetatik garrantzitsuena plangintza berrien martxan jartzea izan da.

Liburu honetan, irakasgaiak funtsean aurrekoak badira ere lizentziatura berriei egokiturik daude. Ekonomia lizentziaturaren "Estatistikarako Sarrera" asignaturak 6 kreditu ditu, hau da, 60 ordu eta Administrazio eta Enpresa Zuzendaritza lizentziaturarenak 3 kreditu, hau da, 30 ordu.

Irakasgai guztiek, Ekonomia lizentziaturako asignaturaren egitaraua betetzen dute.

Administrazio eta Enpresa Zuzendaritza lizentziaturaren egitaraua lehen VI irakasgaiek beteko dute, gainera gai hauetan zenbait frogapen ez da emango. VII, VIII eta IX irakasgaiak "Estatistikarako Sarrera" asignatura gainditzeko, behar ez badira ere, lizentziatura honetan dagoen zenbait berezitasunetan, oso garrantzizkoak izan daitezke.

Bestalde, testuliburu hau Unibertsitateko beste ikastetxe batzuetan, euskaraz azaltzen diren Estatistika asignaturetan, erabilgarria izan daiteke; adibidez, Enpresa Unibertsitate Eskoletan, Gizarte-Graduatuen Eskolan nahiz beste fakultate eta eskoletan. Irakasle eta ikasle askori laguntzeko ilusioz lan egin dugu.

Azken 13 urte hauetan, asignaturaren egitarauetan egindako talde-lana izan da eta meritua ez da gurea bakarrik. Talde hauetan eduki ditugun lankide guztiak biziki eskertu nahi ditugu eta bereziki guztiok oroimenean daukagun J.M. Piris.

Bestalde, euskararen aldetik lagun askoren laguntza eskergaitza izan da, gehienak UEUren ingurukoak direlarik. Eskerrak guztiei eta edizio honetan hainbeste lagundu diguten UEUko Mari Karmen Menika, Nekane Intxaurtza eta Izaro Gurrutxaga.

Egileak
Sarriko, 1996.eko urtarrila

I. EZAUGARRI ESTATISTIKO BAKUNAK, BANAKETAK, TAULAK, ADIERAZPIDE GRAFIKOAK

I.1. ESTATISTIKAREN HISTORIA LABURRA. ESTATISTIKA ETA PROBABILITATEA

I.2. OROKORTASUNAK

I.2.1. Populazioa eta lagina

I.2.2. Unitate estatistiko edo indibiduoak

I.3. EZAUGARRI ESTATISTIKOAK

I.3.1. Ezaugarri estatistiko bakunak

I.3.1.1. Ezaugarri kuantitatiboak

I.3.1.2. Ezaugarri kualitatiboak

I.4. MAIZTASUN-BANAKETA BAKUNAK

I.4.1. Maiztasun-banaketak (absolutuak, erlatiboak eta metatuak)

I.4.2. Taulak

I.5. MAIZTASUN-BANAKETA BAKUNEN ADIERAZPIDE GRAFIKOAK

I.5.1. Barra-diagramak

I.5.2. Histogramak

I.5.3. Maiztasun-poligonoak

I.5.4. Diagrama linealak

. *I.5.5. Beste adierazpide grafikoak*

I.1. ESTATISTIKAREN HISTORIA LABURRA. ESTATISTIKA ETA PROBABILITATEA

Estatistika hitza gutxienez hiru zentzu desberdinetan erabil daiteke.

Esan ohi da:
- "estatistika batzuk" izen arrunt bezala erabiliz edo.
- "..... txosten estatistiko bat....." adjetibotzat hartuz edo.
- ".... Estatistika ikastea.... " izen propio bezala azkenik.

Hitz honen adierazpen orokor, zahar eta ez-espezializatua hauxe da:

MULTZO HANDIEN "Ikasketa" edo "erabilera", MULTZO ZENTZUAN HARTURIK.

Honela, **hasieran** Estatistika artea eta teknika izan zen.

Mundu neolitikoan herriak eta salerosketak gero eta ugariagoak egin zirenez, Estatistikaren beharra sortu zen.

Gizartearen hasiera historikotik praktikatu da beraz. Israeldar, erromatar, txinatar eta inken aparteko inperioarenak, adibide eder batzuk besterik ez dira.

Baina "Cinquecento" italiarrean (XVI. mendean) izena eta bere buruaren kontzientzia hartzen hasi zela dirudi, kontabilitatearekin nahiko lotuta hasi ere.

XVIII. mendetik aurrera argiago eta garbiago agertzen da, beharbada. Estatu-zientziaren alboan hasi zen, gehienbat, Alemaniak emandako bultzada medio.

Etapa honetan, oraindik Estatistika ez da zientzia, eta gertaera aleatorioetatik eta probabilitate-teoriatik urrun zegoen artea zen.

Baina geroztik "estatistikariei" arazoak agertzen hasi zitzaizkien:

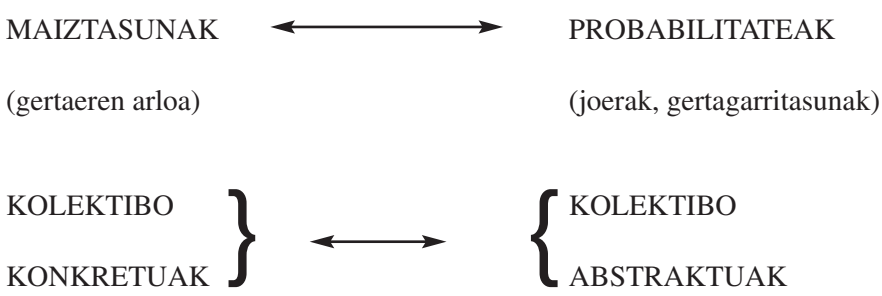
- nola lortu ondorio zientifikoak estatistiketatik?
- nola lortu informazio fidagarriak kolektibo aldakor edo zentsatzeko zail direnetatik?

Britaniar giroan XIX. mendearen amaieran eta XX.aren hasieran, oztopo hauetatik **irteten zen heinean**, Estatistika zientzia bilakatu zen.

Orain Estatistika probabilitate-teoriarekin esentziazko harremanetan dagoen zientzia da.

Honela, hurrengo galderetan ikusiko dugun maiztasun kontzeptutik probabilitate kontzeptura pasatu zen.

Eskematikoki ikusiz:



Fenomenoetatik bere kausetarainoko bidea egiten duen ezagupide honi, INDUKZIOA deritzo Estatistikaren alorrean eta, hementxe, INFERENTZIA ESTATISTIKOA.

Estatistika zientziaren oinarri trinkoan bermatutako estimazioak egiten dira gaur egun, beraien fidagarritasuna ere neurtzen delarik.

Hau dena INFERENTZIA ESTATISTIKOARI dagokio.

Apunte-liburu honetan ikusten diren gaiak Estatistika Deskribatzaileari dagozkio; hau da, oinarrizko Estatistika ikusiko dugu edo Estatistikaren hastapenak, eta halaber, Estatistika Deskribatzaile Anizkoitza edo Datu-Analisia IX. gaian.

Dakusagun, bada, gaian sartu aurretik, ESTATISTIKA DESKRIBATZAILEAREN definizio bat.

ESTATISTIKA DESKRIBATZAILEA, datu-multzo HANDIAK aurkez ditzakeen teknika bat da, multzo hau eskurakoi bihurtuz, beraren egitura gardenduz eta beraren barne-harremanak neurtuz.

I.2. OROKORTASUNAK

I.2.1. Populazioa eta lagina

Ikerketa estatistiko bakoitzean objektu nagusi diren pertsona, ondasun, saioaren emaitza (adibidez: datuaren aurpegia), probintzia baten herri, hauteskundearen alderdi politiko, edota fabrikazio-unitateek osatzen duten multzoari POPULAZIO edo UNIBERTSO deritzogu.

Populazioa ondo definitzeko edozein elementu partikular populaziokoa den ala ez jakin ahal izatea, beharrezkoa zaigu.

Populazioak **bukaezinak** (adb: boltsa bateko erauzketak) ala **bukakorrak** (eskualde bateko enpresa kooperatiboak) izan daitezke.

LAGINA: Lagina azpipopulazio errepresentagarri bat besterik ez da; populazio batez konklusio fidagarriak atera nahi baditugu, ahal dugun azpipopulazio errepresentagarriena aukeratu beharko dugu lagintzat.

LAGIN errepresentagarri hau erauzten duen prozesuari **zorizko laginketa** deritzo.

Lagina erauztean populazioaren elementu guztiei posibilitate berdina ematen badiegu, **zorizko lagin sinplea** deritzogu honela aukeratutako laginari.

Lagina erauzteko beste era eta jokabide asko daude, geruzatuz, mordoketaz... baina hau **LAGINKETA-TEORIAri** dagokio.

Lagina osatzen duten elementuen kopuruari laginaren **tamainua** deritzo.

Sinbolikoki:

Ω = Populazioa

Ω' = Lagina $\quad \Omega' = \{\omega_1, \omega_2, \ldots \omega_N\} \subset \Omega$

N = Laginaren tamainua.

I.2.2. Unitate estatistiko edo indibiduoak

Populazioa edo lagina osatzen duen elementu bakoitzari **unitate estatistiko** edo indibiduo deritzogu.

Indibiduo izena, estatistika deskribatzailearen jatorri demografikoari zor diogu.

I.3. EZAUGARRI ESTATISTIKOAK

Ikerketa estatistiko batean helburua ez dugu populazioa edo lagina exhaustiboki ikertzea, beraren ezaugarri batera edo batzuetara mugatuko gara baizik.

Ikertu nahi dugun ezaugarria bakarra bada, **ezaugarri estatistiko bakuna** deritzo.

Populazioaren bi edo ezaugarri gehiago batera ikertu nahi baditugu, **ezaugarri estatistiko anizkoitza** deritzo.

Ezaugarri estatistikoek (bakun ala anizkoitzek) populazio edo laginen multzoko indibiduo bakoitzean **balio indibidual** bat har dezakete.

Lagineko indibiduo-multzoari, ezaugarriaren **balio-multzo** bat dagokio.

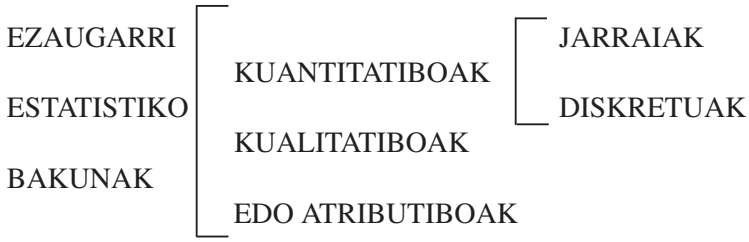
Sinbolikoki:

$\Omega = \{ \omega_1, \omega_2, \ldots\ldots \omega_N \} \longrightarrow \{ x_1, x_2, \ldots\ldots\ldots , x_n \}$

$\Omega = \{ \omega_1, \omega_2, \ldots\ldots \omega_N \} \longrightarrow \{(x_1\ y_1), (x_2\ y_2) \ldots\ldots (x_n\ y_n) \}$

I.3.1. Ezaugarri estatistiko bakunak

Ezaugarri estatistiko bakunen sailkapena bere balio multzo desberdinen arabera:



I.3.1.1. Ezaugarri kuantitatiboak

Ezaugarriaren balio indibidualak populazioan edo laginean **zenbakiak** baldin badira, edota balio-multzoa zenbakizkoa bada, ezaugarri kuantitatiboa deritzo.

Ezaugarriari dagokion balio-multzoa, **multzo diskretua** denean, **ezaugarri diskretua** deritzo.[1]

Balio biren artean edozein balio har dezakenean ezaugarri jarraia deritzo.

I.3.1.2. Ezaugarri kualitatiboak

Ezaugarriak har ditzakeen balio indibidualak kategoria edo atributuak direnean, edota **balio-multzoa** kategoria edo atributuek osatzen dutenean (diskretuak beraz), ezaugarri kualitatiboa deritzo.

Ondo klasifikatu ahal izateko, ezaugarri kualitatiboak hiru baldintza hauek bete behar ditu:

- **Ondo definitua**, hau da, kategoria edo atributu bakoitzak zer ulertarazten duen argiro azaldu behar du.

- **Esklusibotasuna**, hau da, indibiduo bat ezin liteke bi kategoriatan egon.

- **Exhaustibotasuna**, hau da, indibiduo guztiek nonbait klasifikatuak egon behar dute.

ADIBIDEAK: Eskualde bateko enpresa kooperatiboen lagin bat aztertzean:
- Langile-kopurua, ezaugarri kuantitatibo diskretua da;
- Produkzioa, ezaugarri kuantitatibo jarraia da;

1. Balio-multzoko elementuak, puntu isolatuak direnean, balio-multzo diskretu bat dugu.

- Lekutasuna, ezaugarri kualitatiboa da.

I.4. MAIZTASUN-BANAKETA BAKUNAK

I.4.1. Maiztasun-banaketak (absolutuak, erlatiboak eta metatuak)

Ezaugarri diskretuetan:

Suposa dezagun N indibiduoz osatutako lagin bat, non ezaugarri bakun batek "m" balio edo kategoria desberdin hartzen dituen.

Hots:

$$\{ \omega_1, \omega_2, ... \ \omega_N \} \longrightarrow \{ x_1, x_2, ... , x_j ..... \ x_m \}$$
$$\{ \omega_1, \omega_2, ... \ \omega_N \} \qquad \{A,B, ... , J, ... M \}$$

x_j **balioaren MAIZTASUN ABSOLUTUA, sinbolikoki** n_j **ikurraz** adieraziko dugu, eta laginean x_j balioa hartzen duten indibiduoen kopurua da.

Non:

$$\sum_{j=1}^{m} n_j = N$$

x_j **balioaren MAIZTASUN ERLATIBOA, sinbolikoki** f_j **ikurraz** adieraziko dugu, non:

$$f_j = \frac{n_j}{N} \qquad \sum_{j=1}^{m} n_j = N$$

hau da, laginean x_j balioa hartzen duten indibiduoen kopurua erlatiboki kontutan hartuz lortzen den proportzioa da.

Maiztasun erlatiboak portzentaietan ere adierazten dira, batzuetan nahiko komenigarria baita horrela egitea.

Orduan:

$$P_j = \%f_j \cdot 100 \qquad \sum_{j=1}^{m} p_j = 100$$

Laginari dagokion **balio-multzoa ordenatu** ahal badugu (ezaugarria kuantitatiboa bada, beti; bestela ez beti), txikitik handira ordenatuko dugu.

Hots:

$$x_1 \leq x_2 \leq \ldots\ldots \leq x_j \ \ldots\ldots \leq \ x_m$$

Kasu honetan, x_j **balioaren MAIZTASUN ABSOLUTU METATUA,** sinbolikoki N_j , honela definituko dugu:

$$N_j = n_1 + n_2 + \ldots + n_j \qquad N_m = \sum_{j=1}^{m} n_j = N$$

hau da, laginean x_j balioa eta txikiagoak hartzen dituzten indibiduoen kopurua da.

Eta: x_j **balioaren MAIZTASUN ERLATIBO METATUA**, sinbolikoki F_j, honela definituko dugu:

$$F_j = \frac{n_1 + n_2 + \ldots + n_j}{N} = \frac{N_j}{N} = f_1 + \ldots + f_j$$

$$F_m = \sum_{j=1}^{m} f_j = 1$$

hau da, x_j balioa eta txikiagoak hartzen dituzten indibiduoen kopurua N indibiduo guztiekiko kontsideratutako proportzioa (batekotan).

Edozein maiztasun erlatibo bezala, portzentaietan ere adierazten dira.

$$\text{Kasu honetan,} \quad F_m = \sum_{j=1}^{m} p_j = 100$$

I.4.2. Taulak

Datu edo balioen ordenaketa eta elkarketari TABULAZIOA deritzo, eta, azkeneko datu-disposizioari, TAULA ESTATISTIKOA.

Lagin bati dagokion **maiztasun-banaketen taula** (ezaugarri bakun eta diskretua izanik) ondokoa da.

Lagina	Balio--multzoa	MAIZTASUN-BANAKETA			
		M. absolutuak	M. absolutu metatuak	M. erlatiboak	M. erlatibo metatuak
ω_1	x_1 A	n_1	N_1	f_1	F_1
ω_2	x_2 B	n_2	N_2	f_2	F_2
..	.. ..	..	..	..	
ω_j	x_j J	n_j	N_j	f_j	F_j
..	.. ..	..	..	..	
ω_N	x_m M	n_m	N	f_m	1

Ezaugarri jarraietan

Nahiz eta **teorikoki** ezaugarria etengabea izan eta fenomeno askok ezaugarri mota honi erantzun, **praktikan** zehaztasun gehiago edo gutxiago duten tresnez neurtzen direnez gero, diskretu bukaezinak gertatzen zaizkigu.

Hots:

$$\{\omega_1,\omega_2,\ldots,\omega_N\} \xrightarrow{x} \{x_1,x_2,\ldots,x_N\} \subset (a,b) \subset R$$

baina praktikan:

$$\{\omega_1,\omega_2,\ldots,\omega_N\} \xrightarrow{x} \{x_1,x_2,\ldots,x_N..\} \subset (a,b) \subset R$$

Lehen urratsean: ezaugarriari dagokion **balio-multzoa mailatan zatituko dugu**, mailen muturrak ondo definituz.

$$I = (a_0,a_1) \cup (a_1,a_2) \cup \ldots \cup (a_{n-1},a_n)$$

non:

$a \leq a_0 \leq a_1 \leq a_2 \leq .... \leq a_{n-1} \leq a_n \leq b$

Mailen kopurua, tamainu desberdinekoak hartu ala ez..... azterketa bakoitzean erabaki beharreko gauzak dira, baina beti ere mailaketak hiru baldintza hauek bete beharko ditu:

- Ondo definituak izatea
- Esklusibotasuna
- Exhaustibotasuna

Bigarren urratsean: maila bakoitza balio batera mugatuko dugu, **klase--ordezkari** deritzoguna.

Normalki ordezkaria, maila tartearen **erdiko** puntua da, baina goi ala beheko muturrean ere izan liteke.

Honela bada, balio-multzoko maila bakoitza, balio batera mugatzen dugu (ordezkariaren baliora) eta laburpen honetaz informazio estatistikoa galdu arren, hasieran ezaugarri jarraia zena, ezaugarri diskretu bukakor bihurtzen dugu, eta dagokion maiztasun-banaketa kalkulatu ondoren, ondoan dagoen taula bezalako batean adieraz daiteke.

Lagina	Mailak edo klaseak	Klase--ordezkariak	M. absolutuak	M. absolutu metatuak	M. erlatiboak	M. erlatibo metatuak
ω_1	(a_0,a_1)	x_1	n_1	N_1	f_1	F_1
ω_2	(a_1,a_2)	x_2	n_2	N_2	f_2	F_2
..	.. ..	..	..		..	
	(a_{j-1},a_j)	x_j	n_j	N_j	f_j	F_j
..	.. ..	..	..		..	
ω_N	$(a_{m-1},a_{m)}$	x_m	n_m	N	f_m	1
			$\sum^{m} n_j = N$		$\sum^{n} f_j = 1$	

ADIBIDEAK

a) X ezaugarri kualitatiboa da.

Eskualde bateko populazioaren banaketa produkzio-sektoreen arabera:

Sektoreak	Portzentaiak	Portzentaia metatuak
Nekazaritza	70	70
Industrigintza	25	95
Zerbitzuak	5	100

(Kasu honetan, sektore-kategoria bakoitzari dagozkien maiztasun erlatiboak portzentaietan bakarrik jarri ditugu).

b) X ezaugarria kuantitatiboa da.

b-1) **Diskretu-taldekatua**

Enpresen banaketa, langile kopuruaren arabera herri bateko enpresetatik lortutako lagin batean.

Langile kopurua	Enpresa kopurua	
	m. absolutuak	m. abs. metatuak
0-tik 9-ra	3.876	3.876
10-tik 49-ra	2.028	5.904
50-tik 99-ra	976	6.880
100-tik 499-ra	689	7.569
500-tik 999-ra	320	7.889
1.000-tik 4.999-ra	304	8.193
5.000-tik 9.999-ra	295	8.488
10.000-tik aurrera	12	8.500
	N = 8.500	

(Kasu honetan, laginean langile kopuru ezaugarriari zegozkion balio desberdinak asko zirenez gero, taldekatu egin ditugu mailaka edo klaseka).

b-2) **Ezaugarri jarraia**

Klaseak (altuera m.)	Klase--ordezkariak	Portzentaiak	Portzentaia metatuak
1'5 b. gutxiago		1	1
1'5-tik 1'55-era	1'525	7	8
1'55 1'60	1'575	12	20
1'60 1'65	1'625	18	38
1'65 1'70	1'675	30	68
1'70 1'75	1'725	20	88
1'75 1'80	1'775	10	98
1'80 b. gehiago		2	100

Soldaduzkara urte batean doazen gazteen banaketa, beren altueraren arabera:

(Kasu honetan, mailaketan, maila guztiek tamainu berdina dute, aurrenekoa eta azkenekoa kenduz).

I.5. MAIZTASUN-BANAKETA BAKUNEN ADIERAZPIDE GRAFIKOAK

Nahiz eta maiztasun-banaketen taulak informazio guztia hartu (balioak mailetan laburtuak izan ez badira behintzat), grafikoki adierazteak asko laguntzen du ikusarazten eta ulertzen.

Adierazpide grafiko asko erabili izan arren, batzuk bakarrik ikusiko ditugu.

I.5.1. Barra-diagramak

Oso baliagarriak zaizkigu ezaugarriak **diskretuak** ditugunean (kuantitatiboak zein kualitatiboak).

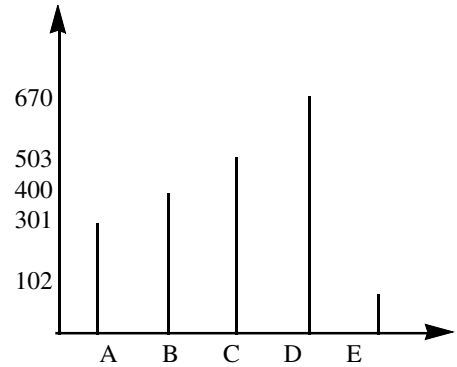
Abzisa-ardatzean balio edo kategoria desberdinak (ahal bada ordenaturik), ezarriko ditugu, eta ordenatu-ardatzean berriz laginari dagozkion maiztasunak (absolutuak, erlatiboak ala metatuak).

ADIBIDEAK

a) Ezaugarri kualitatibo diskretua.

Enpresen banaketa bere kokapenaren arabera:

kokapena	Enpresen kopurua
A arloa	301
B arloa	400
C arloa	503
D arloa	670
E arloa	102
	1.976

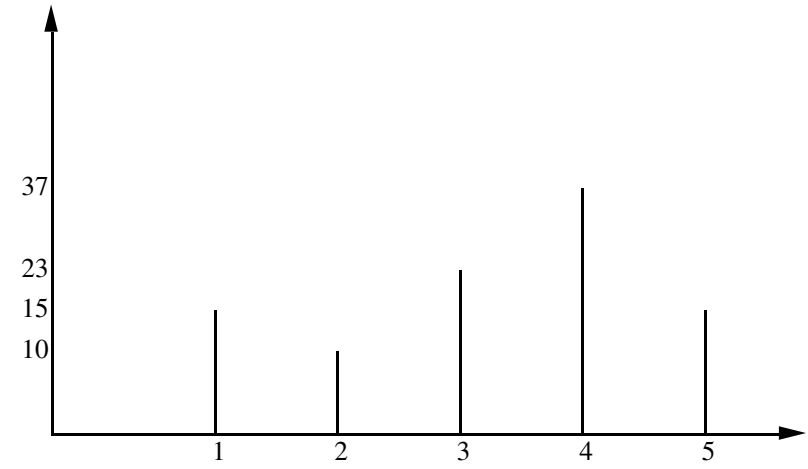


barra-diagrama

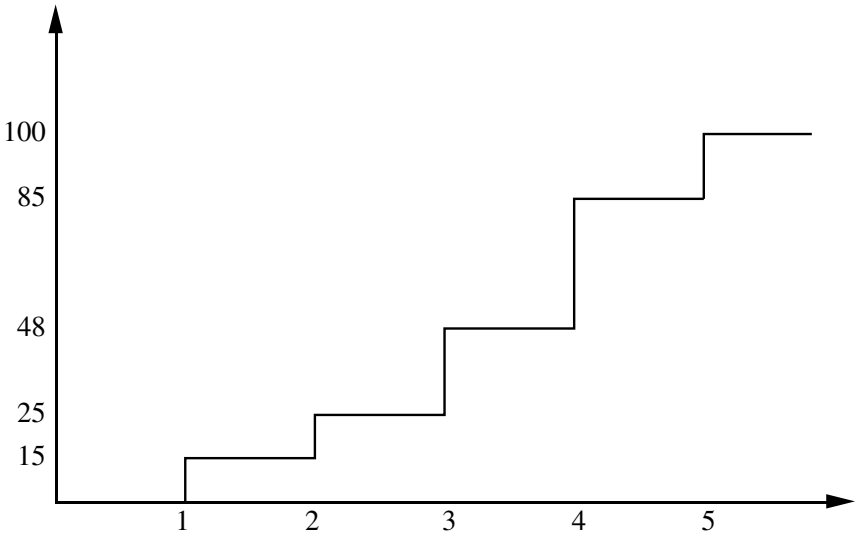
b) Ezaugarri kuantitatibo diskretua.

Loteen banaketa, beren baitan dituzten pieza akastunen kopuruaren arabera:

Txarren kopurua:	Maiztasun erlatiboa: (portzentaietan)	Maiztasun erlatiboa: Portzentaia metatuak
1	15	15
2	10	25
3	23	48
4	37	85
5	15	100



Maiztasun erlatiboen barra-diagrama



Maiztasun metatuen diagrama

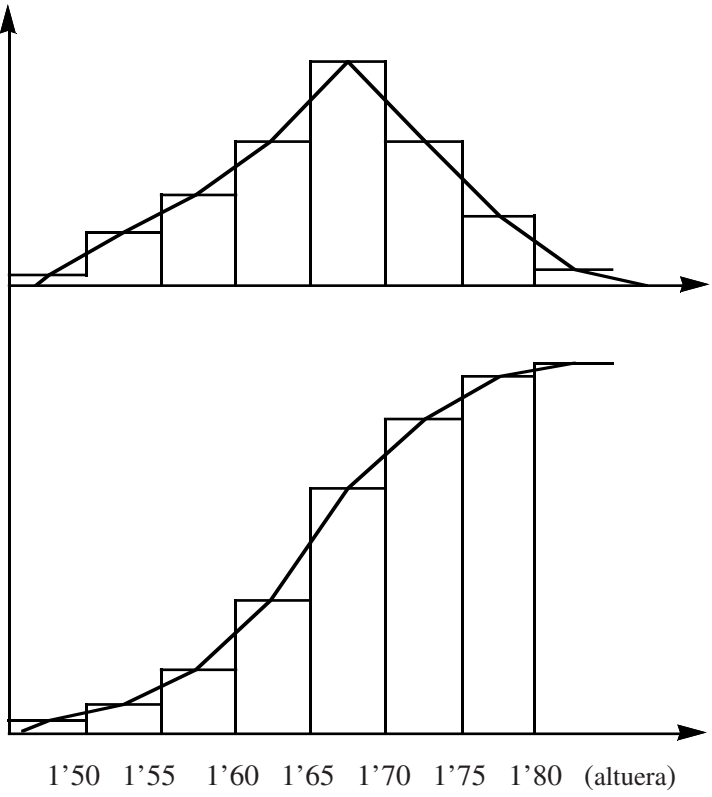
I.5.2. Histogramak

Oso baliagarriak zaizkigu ezaugarriak jarraiak direnean.

Ezaugarriaren balio-multzoa mailetan zatitu ondoren eta maila edo klase bakoitza oinarritzat harturik, dagokion maiztasunaren (nahiz absolutua, erlatiboa edo metatua) **araberako azalera duen laukizuzen bat eraikitzen da**.

Klase dagokion
laukizuzen ela, soilik,
laukizuzen

Ondo oligonoak
ditugu (iku



Histogramak eta maiztasun-poligonoak

Klaseen zabalerak, ordea, desberdinak baldin badira, laukizuzenen altuerak, maiztasun absolutuak edo erlatiboak zabaleraz zatituz, kalkulatuko ditugu.

h_i, n_i eta c_i , i . klasearen altuera, maiztasuna eta zabalera, hurrenez hurren, izanik:

$$h_i = \frac{n_i}{c_i}, \text{ orduan } S_i \text{ azalera: } S_i = c_i \frac{n_i}{c_i} = n_i$$

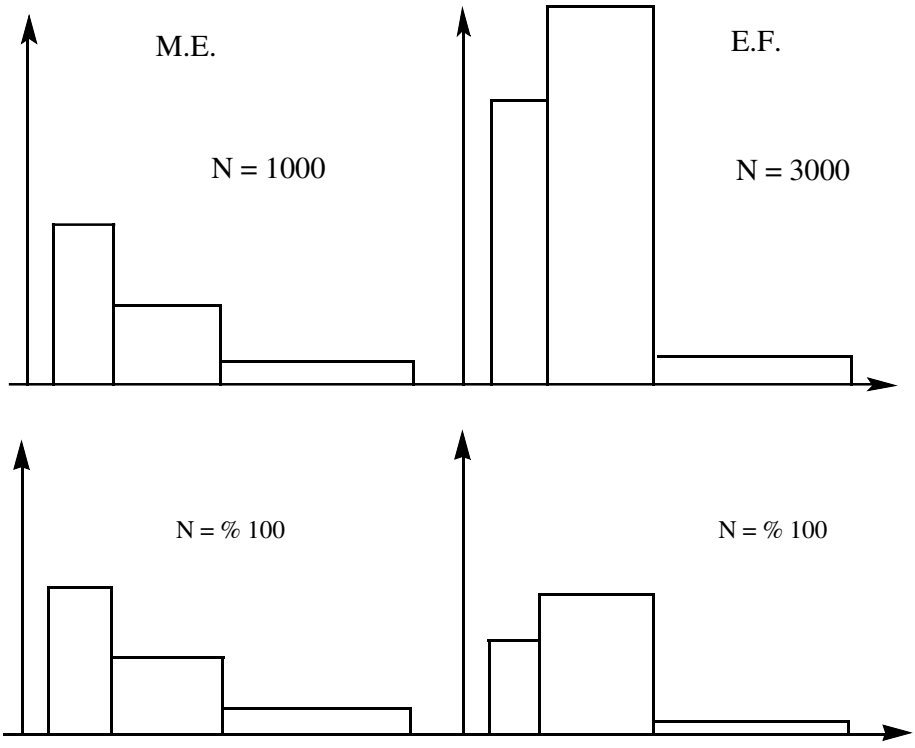
h_i altuera d_i bezala ere izendatzen da eta maiztasun-dentsitate bezala ezagutzen.

Adibidea:

Bi goi-mailako ikastetxetan, ikasleen adinak honela banatzen dira:

Urteak (mailak)	M.E. Ikastetxea	E.F. Ikastetxea
17 - 19	400	750
20 - 25	400	2.000
26 - 37	200	250

Adibidearen histogramak egin ditugu a) kasuan maiztasun absolutuak kontsideratuz eta b) kasuan, berriz, maiztasun erlatiboak portzentaietan.



Dakusagunez, maiztasun erlatiboak (ehunekotan edo ez) adieraztean totalaren ikusketa, N-rena hain zuzen, galtzen da, baina ordea bi banaketak konparagarriago bihurtzen dira.

I.5.3. Maiztasun-poligonoak

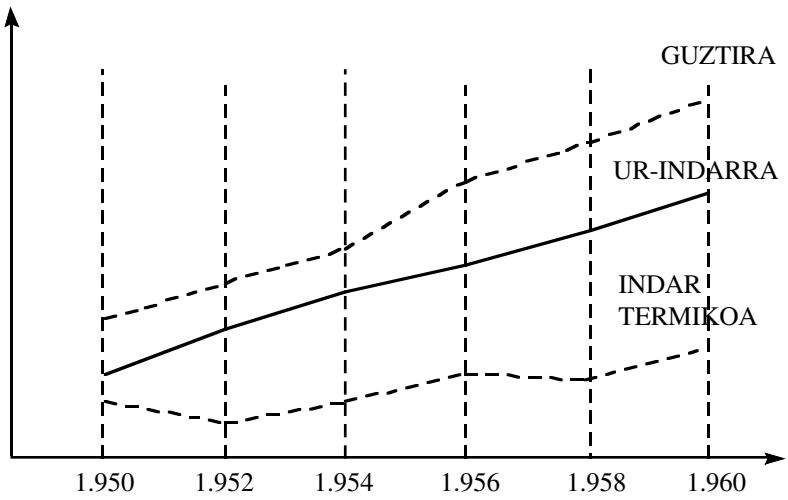
Maiztasun-poligonoa laukizuzenen goiko oinarriko erdiko puntuak marra batez lotuz eraikitzen da.

Honela egin dugu 16. orrialdeko histograman.

I.5.4. Diagrama linealak

Oso erabilgarriak izan ohi dira, balioen aldaketetan denborak izan lezakeen eragina aztertzeko.

Adibidez: Energi produkzioa 1.950etik, 1.960 urterarte:



I.5.5. Beste adierazpide grafikoak

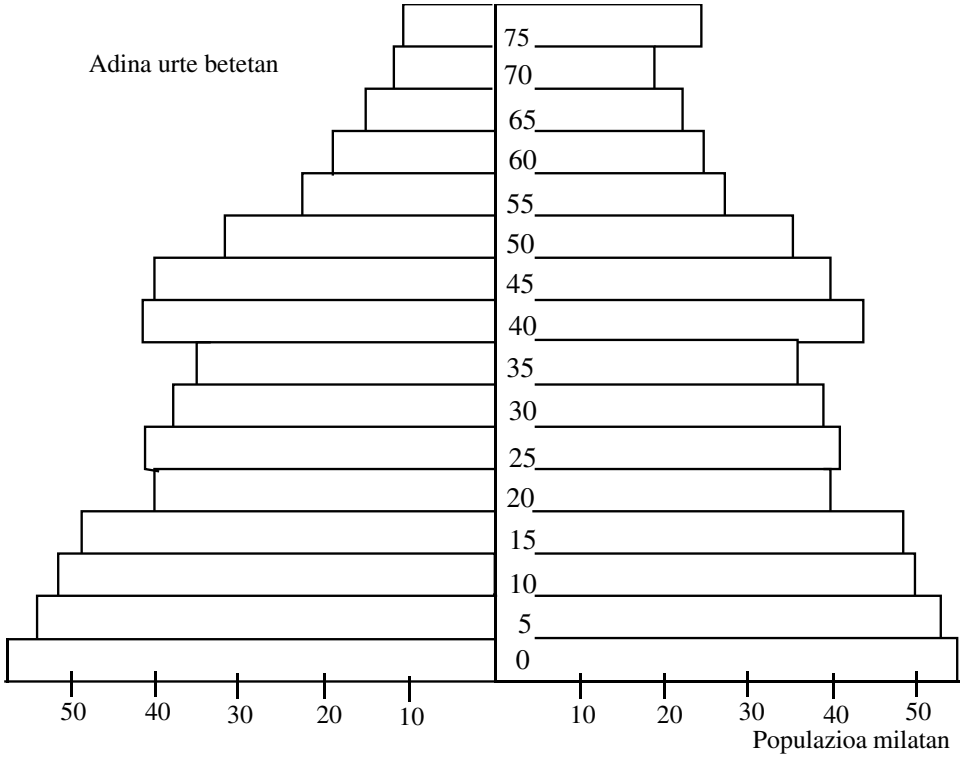
Beste batzuen artean, ikus ditzagun, azkenez, bi hauek:

- POPULAZIO-PIRAMIDEAK.

Histograma bikoitzak dira, eta honela eraikitzen dira: ordenatu-ardatzean adin-taldeak adierazten dira eta abzisa-ardatzean, ezkerrerantz eta eskuinerantz gizonezko eta emakumezkoen maiztasun erlatiboak populazio totalarekiko adierazten dira, dagozkien laukizuzenak eraikiz.

Maiztasun erlatiboak populazio totalarekiko kontsideratzean, gizonezkoen eta emakumezkoen laukizuzenak konpara daitezke adin-talde bakoitzerako.

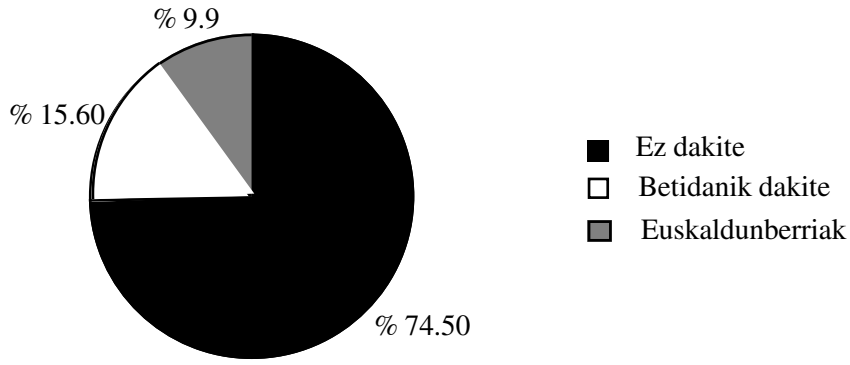
Halaber, maiztasun erlatiboak kontsideratzean, eskala berdinean eraikiz gero, tamainuaren aldetik oso desberdinak diren populazioak konpara daitezke.



Bizkaiko populazio-piramidea 1.975. urtean

- SEKTORE GRAFIKOAK

Grafiko hauek honela eraikitzen dira: zirkunferentziako 360 graduetatik, aldagaien balio ala kategoria bakoitzaren maiztasunari proportzionalki zati bat dagokio.



Euskararen ezaguera Sarrikoko Fakultatean 1979. urtean.

II. EZAUGARRI BAKUNEN BALIO TIPIKOAK

II.1. MOMENTUAK

II.1.1. Momentu arruntak edo jatorriarekikoak

II.1.2. Momentu zentralak edo batezbestekoarekikoak

II.2. MAIZTASUN-BANAKETA BAKUNEN BALIO TIPIKO EDO ESTATISTIKOAK

II.2.1. Zentru-joeraren balio tipikoak

II.2.1.1. Batezbesteko aritmetikoa

II.2.1.2. Batezbesteko geometrikoa

II.2.1.3. Batezbesteko harmonikoa

II.2.1.4. Moda

II.2.1.5. Mediana

II.2.2. Sakabanatzearen balio tipikoak

II.2.2.1. Bariantza

II.2.2.2. Batezbesteko desbidazioa

II.2.2.3. Ibiltartea

II.2.3. Itxuraren balio tipikoak

II.2.3.1. Asimetri koefizientea

II.2.3.1. Kurtosi koefizientea edo zapaltasun/zorroztasun-koefizientea

II.3. KONTZENTRAZIO-NEURKETAK

II.3.1. LORENZ-en kurba

II.3.2. GINI-ren indizea

II.1. MOMENTUAK

Datuek berekin duten informazioa, balio tipiko edo estatistikoen bidez sintetizatzea komeni da.

Noski, laburketa guztiei informazio-galtze bat dagokie; errakuntzarako motibo hau, bada, kontuan izan beharko dugu estatistiko bakoitza aztertzerakoan.

Baina, estatistiko edo balio tipikoak aurkeztu baino lehen, momentuak ikusi behar ditugu, balio tipikoak momentuak baitira edo momentuen bidez definitzen.

Momentuak bi motatakoak izan daitezke:
- arruntak edo jatorriarekikoak.
- zentralak edo batezbestekoarekikoak.

II.1.1. Momentu arruntak edo jatorriarekikoak

Maiztasun-banaketa bakun baten h-ordenako momentu arrunta edo jatorriarekikoa, sinbolikoki a_h ikurraz adieraziko da eta honela definituko:
non:

N = Laginaren edo populazioaren indibiduoen kopurua.
x_i = Ezaugarriaren i. balioa.
m = Ezaugarriak hartzen dituen balio desberdinak; $m \leq N$ izango da.

$$a_h = \frac{\sum_{i=1}^{N} (x_i)^h}{N} = \frac{\sum_{i=1}^{m} (x_i)^h \; n_i}{N}$$

Momentu arrunten bi berezitasun ikusiko ditugu:

1) O-ordenako momentuak 1 balioa hartzen du:

Hots:

$$a_0 = \frac{\sum_{i=1}^{m} (x_i)^0 \; n_i}{N} = 1$$

2) 1- ordenako edo lehenengo ordenako momentua batezbesteko aritmetikoa da:

Hots:

$$a_1 = \frac{\sum_{i=1}^{m} x_i \; n_i}{N} = \bar{x}$$

eta normalki $\bar{x}$ ikurraz adierazten da.

Ikusten ari garen momentu arruntek nahiz ondoren ikusiko ditugunek, ez dute berekiko esangurarik, bide batez balio tipiko direnean izan ezik, adibidez, batezbesteko aritmetikoa, bariantza, e.a.

Halere, Estatistikan eragile matematiko bezala sartzen dira, teoriaren garapenean oso erabilgarriak baitira.

II.1.2. Momentu zentratuak edo batezbestekoarekikoak

Maiztasun-banaketa bakun baten h-ordenako momentu zentratua edo batezbestekoarekikoa, sinbolikoki m_h ikurraz adieraziko da eta honela definituko:

$$m_h = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^h}{N} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^h \; n_i}{N}$$

Eragiketak eginez, ikus dezagun nola kalkulatzen diren momentu zentratuak momentu arrunten bidez:

$$m_h = \frac{1}{N} \sum^{N} (x_i - \bar{x})^h \; n_i =$$

$$= \frac{1}{N} \sum^{m} \left(\binom{h}{0} x_i^h - \binom{h}{1} x_i^{h-1} \bar{x} + \binom{h}{2} x_i^{h-2} \bar{x}^2 - \ldots\ldots + \right.$$

$$\left. + (-1)^{h-1} \binom{h}{h-1} x_i \bar{x}^{h-1} + (-1)^{h-1} \binom{h}{h} \bar{x}^h \right) n_i =$$

$$= \binom{h}{0} a_h - \binom{h}{1} a_{h-1} \bar{x} + \binom{h}{2} a_{h-2} \bar{x}^2 - \ldots + (-1)^{h-1} \; h\bar{x}^h + (-1)^h \bar{x}^h$$

Hortik:

$$m_2 = a_2 - \bar{x}^2$$

$$m_3 = a_3 - 3a_2\bar{x} + 2\bar{x}^3$$

$$m_4 = a_4 - 4a_3\bar{x} + 6a_2\bar{x}^3 - 3\bar{x}^4$$

Momentu hauen artean bigarren ordenakoa (m_2), BARIANTZA delakoa da garrantzitsuena.

$$m_2 = \frac{\sum (x_i - \bar{x})^2 \; n_i}{N} = a_2 - \bar{x}^2 = \frac{\sum x_i^2 \; n_i}{N} - \left(\frac{\sum x_i \; n_i}{N} \right)^2$$

Eta azkenez, momentuak edozein punturekiko izan daitezke:

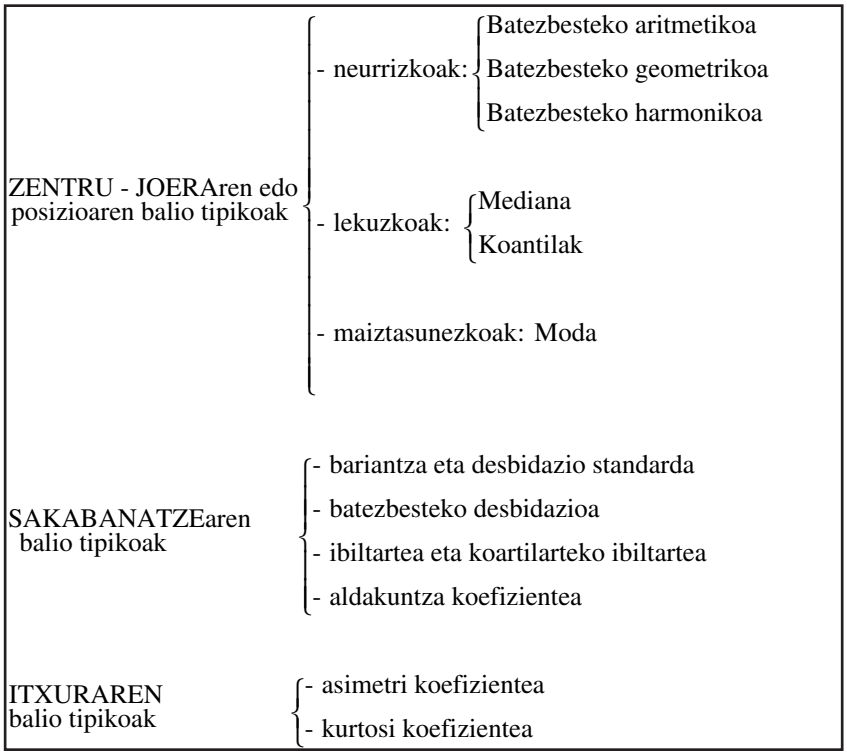
$$a_h, \; x_0 = \frac{\sum^{N} (x_i - x_0)^h}{N} = \frac{\sum^{m} (x_i - x_0)^h \; n_i}{N}$$

$x_0 = 0$ baldin bada, momentu arruntak izango dira.

eta $x_0 = \bar{x}$ bada, zentratuak.

II.2. MAIZTASUN-BANAKETA BAKUNEN BALIO TIPIKO EDO ESTATISTIKOAK

Organigrama batean azalduko ditugu lehenik banaketa bakunetan erabili ahal diren estatistiko edo balio tipikoak, eta gero erabilienak banaka ikusten joango gara.



II.2.1. Zentru-joeraren edo Posizioaren balio tipikoak

Lagin batean ezaugarri bati dagozkion balioak ordenatzen baditugu, zentruan kokatzen diren balioak erabiltzen dira datu guztiak errepresentatzeko. Posizio--neurriak dira.

Balio batean datu-multzo baten ideia ematea argiro ikusten dugu, sarritan gehiegizko sinplifikazioa izanik.

Horregatik, zentruko estatistikoak beste estatistiko batzuekin (sakabanatzearenak..) osatzen dira, aztertzen ari garen populazio edo laginaren ideia errealagoa eduki nahi baldin badugu.

II.2.1.1. Batezbesteko aritmetikoa

Lagin batean $x_1, x_2, ..., x_i, ..., x_m$, zenbakizko balio-multzo bati dagozkion maiztasun absolutuak hurrenez hurren $n_1, n_2, ..., n_i, ..., n_m$ baldin badira, beraren batezbesteko aritmetikoa, $\bar{x}$, honako hau da:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{m} x_i \; n_i \qquad \text{non} \qquad \sum_{i=1}^{m} n_i = N$$

Maiztasun erlatiboak, $f_1 , f_2 , ... , f_m$, kontutan hartuz

$$\bar{x} = \sum_{i=1}^{m} x_i \; f_i$$

Ezaugarriak hartzen dituen balio guztiak desberdinak balira, maiztasun absolutuek 1 balioa hartuko lukete, hau da, $n_1 = n_2 = ... = n_N = 1$ eta orduan:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Maiztasun erlatiboak masak bezala kontsideratzen baditugu, batezbesteko aritmetikoak **grabitate-zentrua** finkatuko du. Hau da, barra baten gainean eta $x_1 , x_2 , ... x_m$ puntuetan maiztasunak bezalako pisuak kokatzen baditugu sistemak oreka lor dezan, esekitze-puntuak $\bar{x}$ puntuan izan beharko du.

PROPIETATEAK:

1.- Edozein maiztasun-banaketatan, bere maiztasunaz biderkatuz, batezbestekoarekiko desbideratzeen batukaria zero egiten da.

Hots:

$$\sum_{i=1}^{m} (x_i - \bar{x}) \; n_i = 0$$

Frogapena:

$$\sum_{i=1}^{m}(x_i - \bar{x})\ n_i = \sum_{i=1}^{m} x_i\ n_i - \bar{x}\ \sum_{i=1}^{m} n_i = N\ \bar{x} - N\ \bar{x} = 0$$

2.- Ezaugarri edo aldagai baten balio guztiak h parametroaz biderkatzen baditugu, batezbesteko aritmetikoa ere h-z biderkatuta edukiko dugu.

Hots:

$$\bar{x} = \frac{1}{N}\ \sum_{i=1}^{m} x_i\ n_i \quad \text{izanez}$$

x_i balio bakoitza h x_i balioaz ordezkatzen badugu:

$$\frac{1}{N}\ \sum_{i=1}^{m} h\ x_i\ n_i = h\frac{1}{N}\ \sum_{i=1}^{m} x_i\ n_i = h\ \bar{x}$$

3.- Aldagai bi edo gehiagoren batura den aldagai baten batezbesteko aritmetikoa, batezbesteko aritmetikoen batura da.

Hau da: $t = x + y + z$ baldin bada.

$\bar{t} = \bar{x} + \bar{y} + \bar{z}$ izango da.

Suposa dezagun oposizio-lehiaketa bat hiru azterketaz osatzen dela eta i. lehiakideak x_i, y_i, z_i puntuak ateratzen dituela, puntuazio totala $t_i = x_i + y_i + z_i$ izanez; N lehiakideren puntuazioen batezbesteko aritmetikoa ondokoa izango da:

$$\bar{t} = \frac{\sum_{i=1}^{N} t_i}{N} = \frac{\sum_{i=1}^{N}(x_i + y_i + z_i)}{N} =$$

$$= \frac{\sum_{i=1}^{N} x_i}{N} + \frac{\sum_{i=1}^{N} y_i}{N} + \frac{\sum_{i=1}^{N} z_i}{N} = \bar{x} + \bar{y} + \bar{z}$$

BATEZBESTEKO ARITMETIKO PONDERATUA

Batzuetan batezbesteko aritmetikoa ateratzean, aldagaien garrantzi erlatiboa desberdina dela ohartzen gara.

Suposa dezagun hiru sagar mota saltzen direla 60, 80 eta 90 pezeta/kilo prezioetan hurrenez hurren.

Batezbesteko prezioa honela atera daiteke:

$\bar{x} = \dfrac{60 + 80 + 90}{3} =$ 76'6, hau da, batezbesteko aritmetiko sinplea eginik.

Baina mota bakoitzarentzat saltzen diren kiloak, hurrenez hurren, hauek baldin badira: 125, 72, 3

Batezbesteko ponderatua, zera izango da:

$$\bar{x}_{(pon)} = \frac{60 \cdot 125 + 80 \cdot 72 + 90 \cdot 3}{125 + 72 + 3} = 67'6$$

Ikusten dugunez, batezbesteko aritmetiko ponderatua egitean, aldagaiaren balio bakoitza zenbaki batez biderkatzen dugu, balio-multzo barnean daukan garrantziaren arabera.

Zenbaki hauek pisuak edo ponderazioak dira eta sinbolikoki w_i deituko ditugu.

Orduan $\bar{x}_{(pon)} = \dfrac{\sum x_i w_i}{\sum w_i}$ edota $\bar{x}_{(pon)} = \sum x_i w_i$ $\sum w_i = 1$ denean

Oharra: kontutan hartu behar da, bien formulak baliokideak badira ere ponderazioak ez direla maiztasunak. Hala ere, maiztasunak ponderazioak dira aldagaiaren balio desberdinek duten garrantzia maiztasunen araberakoa izango baita.

Beste adibide bat:

Ondoko taulan Hego Euskal Herriko lau herrialdeen kasuan hileroko gastuak/familia datuak dauzkagu.

Hileroko gastuak / familia, 1.968. urtean

	Familiak	Gastuak (mila pezetatan)
Araba	24	4,4
Gipuzkoa	53	5,3
Nafarroa	57	5,1
Bizkaia	84	5,1

Batazbesteko aritmetikoa: $\bar{x} = \frac{19,9}{4} = 4,975$

Baina, familien kopurua kontutan hartuz:

$$\bar{x}_{(pon)} = \frac{4,4 \cdot 24 + 5,3 \cdot 53 + 5,1 \cdot 57 + 5,1 \cdot 84}{24 + 53 + 57 + 84} = 5,07$$

Noski, batezbesteko ponderatuen bidez errealitatearen azterketari gehiago hurbiltzen gatzaizkio, baina normalki datu guztiak ezagutzen ez ditugunez, batezbesteko aritmetiko sinplea egiten dugu.

II.2.1.2. Batezbesteko geometrikoa:

Ondoko berdintasunaren bidez adierazten dugu:

$$G = \sqrt[N]{x_1^{n_1} \cdot x_x^{n_2} \ldots\ldots\ldots x_m^{n_m}} = \sqrt[N]{\prod_{i=1}^{m} x_i^{n_i}}$$

Arrazoi edo ehunekoen batezbesteko bat ateratzeko, batezbesteko aritmetikoa baino aproposagoa da eta ondoko propietatea betetzen du:

$$\text{Log } G = \frac{1}{N} \log \prod_{i=1}^{m} x_i^{n_i} = \frac{1}{N} \sum_{i=1}^{m} (\log x_i) n_i$$

Hots:

Batezbesteko geometrikoaren logaritmoa, aldagaiaren balioen logaritmoen batezbesteko aritmetikoa da.

II.2.1.3. Batezbesteko harmonikoa

Ondoko berdintasunaren bidez adierazten dugu

$$H = \frac{N}{\sum_{i=1}^{m} \frac{n_i}{x_i}} \qquad \text{non} \qquad N = \sum_{i=1}^{m} n_i$$

Ohar daiteke batezbesteko harmonikoa zera dela, x_i balioen alderantzizko balioen batezbesteko aritmetikoaren alderantzizko balioa.

Batezbesteko abiaduren batezbesteko bat ateratzeko, batezbesteko harmonikoa da egokiena.

Ikusitako batezbestekoak eta beste batzuk: batezbesteko koadratikoa, batezbesteko kubikoa, e.a. sar daitezke k ordenako batezbestekoaren kontzeptuan.

II.2.1.4. Moda

Maiztasun-banaketa batean gehien errepikatzen den aldagai edo ezaugarriaren balioari, hau da, maiztasunik handiena duenari, moda eta balio modala deritzo.

Hots: $M_{(k)} = \sqrt[k]{\sum x_i^k \ f_i}$

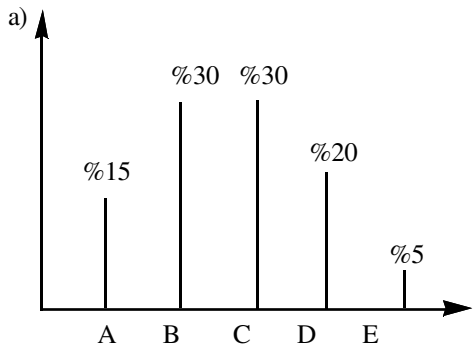
$x_1 , x_2 , ... x_i$ ..... x_m edo A, B, ... , I , ... M

balio edo kategorien multzo bati dagozkien maiztasun erlatiboak hurrenez hurren f_1 , f_2 ,...., f_m baldin badira, maiztasun erlatibo gehien duen x_i balioari edo I kategoriari "moda" M_0 deritzo.

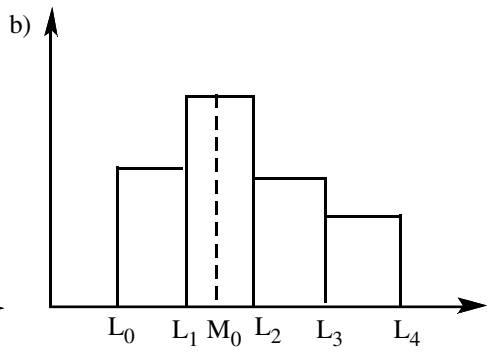
Hots: $M_0 = x_i$ edo $M_0 = I$, n_i eta f_i maiztasun handienak izanik.

Banaketa, modagabea izan liteke $f_1 = f_2 = ... = f_m$, edo moda bat baino gehiago ere eduki dezake, eta honen arabera, “bimodala”, “hirumodala”... orokorki moda-anitza izango da.

ADIBIDEAK



Kasu honetan banaketa bimodala dugu, B eta C arloak modak direlarik.



L_1 , L_2 klasea da kasu honetan moda, klase honen ordezkaria, M_0, har daiteke modatzat.

Banaketa jarraia edo klasetan taldekatua bada, dakigunez, maiztasun handiena, M_0, duen klasean kokatuko da moda, eta klase-ordezkaria modatzat hartzen ez badugu, gutxi gora-beherako kalkuluak egin daitezke.

Klase edo tarte modala L_{i-1}, L_i izanez, balio modala estimatzeko bi irizpide desberdinetan oinarritutako formulak proposatuko ditugu.

1) Altueren alderaketaren irizpidea:

$$M_{(k)} = L_{i-1} + \frac{d_{i-1}}{d_{i-1} + d_{i+1}} \cdot c_i$$

non:

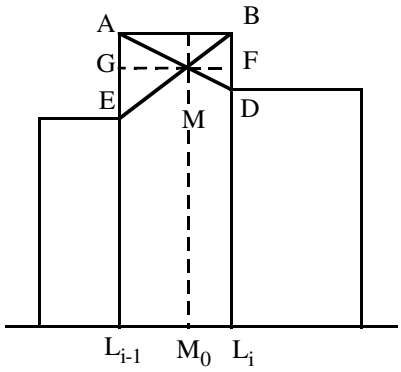
c_i = klaseen zabalera

$d_{i-1} = h_i - h_{i+1}$

$d_{i+1} = n_i - n_{i+1}$

h = i. klaseen altuera

Frogapena grafikoa



AME eta BMD triangeluak antzekoak dira, hau da, beraien altuerak oinarriekiko proportzionalak dira.

Hots:

$$\frac{MG}{MF} = \frac{AE}{BD}$$

Edo ondoko proportzio baliokidea:

$$\frac{MG}{MG + MF} = \frac{AE}{AE + BD} \quad \text{hau da,} \quad \frac{MG}{c_i} = \frac{d_{i-1}}{d_{i-1} + d_{i+1}}$$

eta, $M_0 = L_{i-1} + MG$ izatean, proposatu dugun formula frogatuta gelditzen da.

Klaseak zabalera berdinekoak balira, altueren alderaketa, zuzenki, maiztasunen alderaketa izango da.

Hau da:

$d_{i-1} = n_i - n_{i-1}$

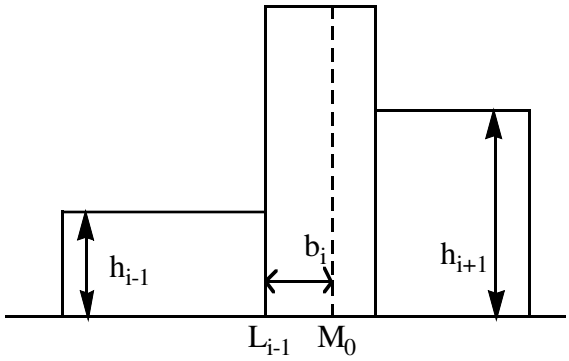
$d_{i+1} = n_i - n_{i+1}$

2) Alderantzizko banaketa proportzionalaren irizpidea:

$$M_0 = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot c_i$$

non: h_{i-1} = aurreko klasearen altuera
h_{i+1} = hurrengo klasearen altuera

Ondoko klaseen altuerekiko alderantzizko banaketa proportzionala egitean, aurreko formula lortzen dugu.



Hots:

$$\frac{b_1}{h_{i+1}} = \frac{c_i - b_i}{h_{i-1}}$$

edo aurrekariak eta atzekariak batzean:

$$\frac{b_i}{h_{i+1}} = \frac{c_i}{h_{i+1} + h_{i-1}}$$

eta, $M_0 = L_{i-1} + b_i$ izatean, proposatu dugun formula frogatuta gelditzen da.

Klaseak zabalera berdinekoak balira, hau da, $c_i = c$
c zabaleraz biderkatzen eta zatitzen badugu eta $n_i = c \cdot h_i$ dela kontutan hartuz, ondoko hau izango da formula:

$$M_0 = L_{i-1} + \frac{h_{i+1} \cdot c}{(h_{i+1} + h_{i-1})c} c = L_{i-1} + \frac{n_{i+1}}{h_{i+1} + h_{i-1}} c$$

II.2.1.5. Mediana eta koantilak

Mediana banaketaren balioak bi zati berdinetan zatitzen dituen lekuzko estatistikorik garrantzitsuena da.

Balio-multzoa ordenatuz gero, maiztasun absolutuen histograma azalera berdineko bi zatitan ebakitzen duen balioari mediana deritzo, eta hori zehazki betetzen duenik ez badago, aurrean dagoena hartzen dugu batzuetan medianatzat.

Ondoko baldintza betetzen duen balio bat da:

$$M_e = x_i \ / \ N_{i-1} < \frac{N}{2} \ \wedge \ N_i > \frac{N}{2}$$

Hala ere, banaketa diskretua bada eta indibiduoen kopurua bakoitia, aurreko baldintza bete egingo da, baina ez beste edozein kasutan.

Banaketa diskretua eta N bikoiti baten aurrean bagaude.

$$M_e = \frac{x_i + x_{i+1}}{2} \Big/ N_i = \frac{N}{2}$$

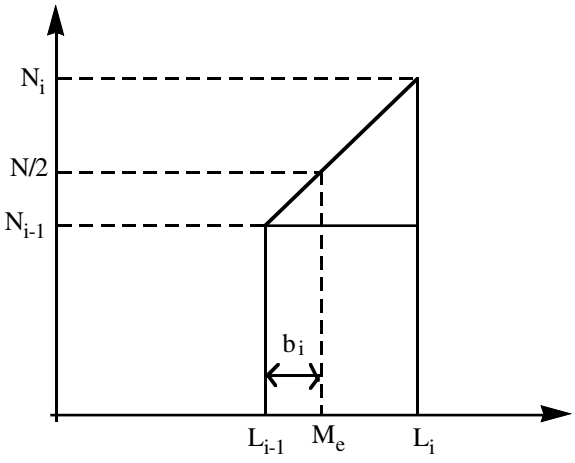
Hots: $M_e = \dfrac{\text{erdiko balioak}}{2}$

<u>Banaketa taldekatuetan eta jarraietan</u>: N/2 balioa duen tartea finkatuz gero eta tartean n_i balioak linealki gehitzen direla suposatuz kalkulatuko dugu M_e.

Orduan: (L_{i-1}, L_i) M_e duen tartea izanez:

$$M_e = L_{i-1} + b_i$$

Irudian dauden bi triangeluak antzekoak izanez, zera idatz daiteke:



$$\frac{N_i - N_{i-1}}{c_i} = \frac{\frac{N}{2} - N_{i-1}}{b_i}$$

$$\text{eta: } b_i = \frac{N/2 - N_{i-1}}{N_i - N_{i-1}}\ c_i$$

Oharra: maiztasun erlatiboak erabiliz, era berdinean kalkula daiteke M_e; kasu horretan,

$$\frac{N}{2} = \frac{1}{2}$$

KOANTILAK

Koantilak beste lekuzko neurriak dira. Mota desberdinetakoak dira: koartilak, dezilak eta zentilak.

Koartilak, dezilak eta zentilak medianaren antzera definitzen dira, baina bi parte hartu ordez lau (koartilak), hamar (dezilak) edo ehun (zentilak) parte hartuz.

Koartilak q_1, q_2, q_3 izango dira; (q_1, q_3) tartean banaketaren erdia daukagu, justu erdiko balioak, hau da, bazterreko balioak kenduz.

Dezilak d_1 ..... d_9 izango dira eta zentilak, c_1 .... c_{99} .

Argiro ikusten da:
$q_2 = d_5 = c_{50} = M_e$ dela.

OHARRAK:

Erdiko baliorik garrantzitsuenak, batezbesteko aritmetikoa, moda eta mediana dira eta bereziki batezbesteko aritmetikoa, guztiak aldagaia edo ezaugarria neurtua dagoen unitateetan neurtuak direlarik.

Ondoko erlazio hau betetzen da:

$$M_0 \simeq 3\ M_e - 2\ \bar{x}$$

Eta gutxi gora-behera hiruretako edozein, besteen bidez kalkula daiteke.

II.2.2. Sakabanatzearen balio tipikoak

Datu-multzoaren ideia orokorra ematen ziguten zentrurako joeraren balio

tipikoek. Ideia hau osatzeko, zentrurako joeraren balioen inguruan datu-multzoa guztiz bilduta edo oso sakabanatuta dagoen adierazten diguten estatistiko batzuk ere behar ditugu.

Batezbesteko aritmetikoaren inguruan datuek duten sakabanatze-neurri bezala, estatistiko garrantzitsuenak hauek dira: Bariantza, Desbidazio Standarda eta Batezbesteko Desbidazioa.

II.2.2.1. Bariantza

Batezbestekoarekiko sakabanatze-neurririk garrantzitsuena da bariantza eta bera erabiliz lortzen ditugu desbidazio standarda eta aldakuntza koefizientea.

$\{x_1 , x_2 , ... , x_i .... x_m\}$ zenbakizko balio-multzo bati dagozkien maiztasun absolutuak hurrenez hurren $n_1 , n_2, ... n_i .... n_m$ baldin badira, bariantza, S_x^2, deritzona hau da:

$$S_x^2 = \frac{1}{N}\sum_{i=1}^{m}(x_i - \overline{x})^2 n_i \quad \text{edota} \quad S_x^2 = \sum_{i=1}^{m}(x_i - \overline{x})^2 f_i$$

Bariantza neurri-unitate karratuetan daukagula erraz ikusten dugu formulan.

Orduan, sakabanatze-neurri bezala bariantzaren erro karratua hartzea egokiagoa da, **Desbidazio Tipikoa edo Standarda**, hain zuzen.

$$S_x = \sqrt{\frac{1}{N}\sum_{i=1}^{m}(x_i - \overline{x})^2 n_i} = \sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2 f_i} = \sqrt{a_2 - \overline{x}^2}$$

eta ezaugarria neurtuta dagoen unitateetan, sakabanatze-neurketa daukagu.

Askotan, estatistika inferentzialean estimatzaile bezala dituzten propietateengatik, S_x^{*2} Quasi-bariantza eta S_x^* Quasi-desbidazio standarda erabiltzen dira.

$$S_x^{*2} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2 n_i}{N-1} \quad \text{eta} \quad S_x^* = \sqrt{\frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{N-1}}$$

Konputagailuko programa-paketeek S_x^{*2} eta S_x^* ematen dizkigute, Bariantza eta Desbidazio Standarda deitzen badituzte ere.

Desbidazio tipikoak batezbestekoarekiko suposatzen duen batekoa (edo ehunekoa) **aldakuntza koefizientea** deitzen da.

g_0 ikurraz adieraziko dugu sinbolikoki, non: $g_0 = \frac{S_x}{\bar{x}}$

$\bar{x}$ batezbestekoa zerora hurbiltzen bada, g_0 horrek ez du zentzu askorik (infiniturantz jotzen baitu kasu honetan).

Bariantzak garrantzizko propietate bat betetzen du, bigarren ordenako momentu zentralik txikiena baita, hain zuzen.

Hots:

$$a_2,\ x_0 = \frac{1}{N}\sum_{i=1}^{m}(x_i - \bar{x}_0)^2\ n_i =$$

$$= \frac{1}{N}\sum_{i=1}^{m}\Big((x_i - \bar{x}) + (\bar{x} - x_0)\Big)^2\ n_i =$$

$$= \frac{1}{N}\sum_{i=1}^{m}\Big((x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - x_0) + (\bar{x} - x_0)^2\Big)\ n_i =$$

$$= \frac{1}{N}\sum_{i=1}^{m}(x_i - \bar{x})^2 n_i + 2(\bar{x} - x_0)\ \sum_{i=1}^{m}(x_i - \bar{x})\frac{n_i}{N} +$$

$$+ \frac{1}{N}\ (\bar{x} - x_0)^2\ \sum_{i=1}^{m} n_i = S_x^2 + (\bar{x} - x_0)^2$$

Eta argiro dagoenez, $x_0 = \bar{x}$ baldin bada, txikiena izango da, beste edozein kasutan bigarren batugaia positiboa baita.

II.2.2.2. Batezbesteko desbidazioa

Balio absolutuak harturik (konpentsaziorik izan ez dadin), batezbesteko batekiko balioek duten desbidazioen batezbesteko aritmetikoari, “batezbesteko desbidazioa” deritzo.

$$\text{Hau da: } D = \frac{1}{N} \sum_{i=1}^{m} \left| x_i - \text{batazbestekoa} \right| n_i$$

a) Aukeratutako batezbestekoa, batezbesteko aritmetikoa bada, batezbesteko desbidazioa hauxe izango da:

$$D_{\bar{x}} = \frac{1}{N} \sum_{i=1}^{m} \left| x_i - \bar{x} \right| n_i$$

Neurri honek indibiduoen homogenotasuna adierazten digu, hau da, desbidazioak txikiak badira, batezbesteko desbidazioa txikia izango da eta, alderantziz, handiak badira, batezbesteko desbidazioa handia izango da.

b) Aukeratutako batezbestekoa mediana bada, batezbesteko desbidazioa hauxe izango da:

$$D_{M_e} = \frac{1}{N} \sum_{i=1}^{m} \left| x_i - M_e \right| n_i$$

Batezbestekoa mediana denean, batezbesteko desbidaziorik txikiena daukagu.

II.2.2.3. Ibiltartea

Ezaugarri zenbakizko batek hartzen dituen balio maximo eta minimoaren arteko diferentziari ibiltarte deritzogu.

Hots:

$x_0 \leq x_1 \leq \ldots\ldots\ldots \leq x_m$ izanik

$R = x_m - x_0$

Koartil arteko ibiltarteak ere kontsidera daitezke, adibidez, $q_3 - q_1$ hirugarren eta lehenengoen artekoena hain zuzen ere, edota, dezil arteko ibiltarteak e.a.

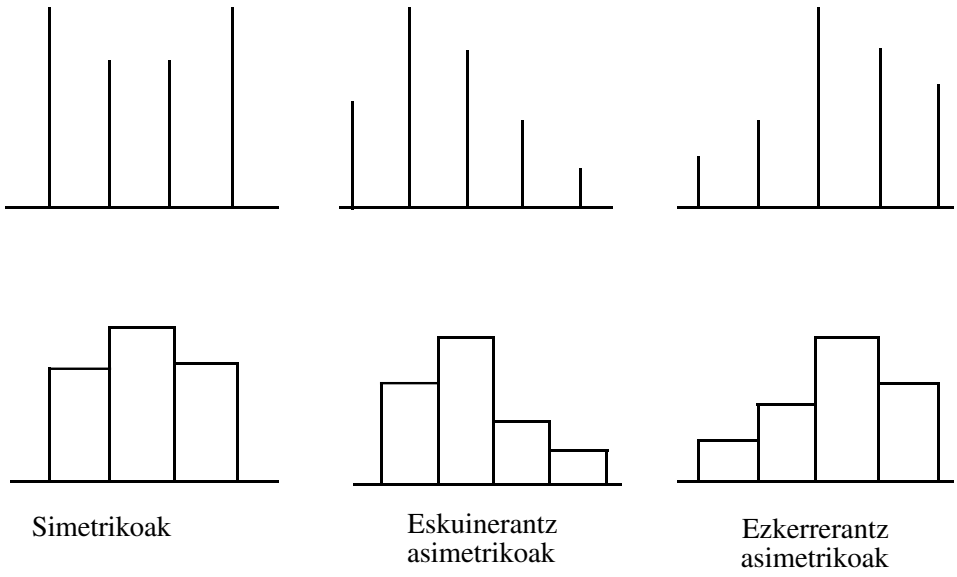
II.2.3. Itxuraren balio tipikoak

Aztertzen ari garen laginaren ezaugarria eta berari dagokion datu-multzoa, zentrurako joeraren eta sakabanatze-balioen bidez nahiko argi geratu bada ere, itxuraren balio tipikoak aztertzean gehiago osa daiteke; hauen artean asimetri koefizientea eta kurtosi koefizientea ikusiko ditugu.

II.2.3.1. Asimetri koefizientea

Maiztasun-banaketaren grafikoa simetrikoa denean, banaketa simetrikoa dela esango dugu, hau da, barra-diagrama edo histograma egitean simetrikoa bada, banaketaren elementuak batezbestekoarekiko bi aldeetan berdin kokatzen direla esan nahi du.

Aurrekoa betetzen ez bada, banaketa asimetrikoa izango da.



Banaketa bat moda batekoa eta simetrikoa baldin bada, batezbesteko aritmetikoa, moda eta mediana balio berean daude.

Hots: $\bar{x} = M_0 = M_e$

Asimetri koefizientea g_1 ikurraz adieraziko dugu sinbolikoki, non:

$$g_1 = \frac{m_3}{S_x^3}$$

$$m_3 = \sum_{i=1}^{m} \left(x_i - \bar{x}\right)^3 \frac{n_i}{N}$$ izanik, maiztasunen balioak $\bar{x}$ balioarekiko

simetrikoki kokatuta badaude, $m_3 = 0$ izango da; baina maiztasunen balioak eskuinaldean **handiagoak** badira, hau da, $x_i > \bar{x}$ denean, orduan $m_3 > 0$ izango da, eta, alderantziz, ezkerraldean handiagoak badira, hau da, $x_i < \bar{x}$ denean, orduan $m_3 < 0$.

Ikusten dugu, bada, m_3 asimetriaren balio adierazle bezala har dezakegula, baina neurri-unitateak ber hiru izanez zuzenkiago.

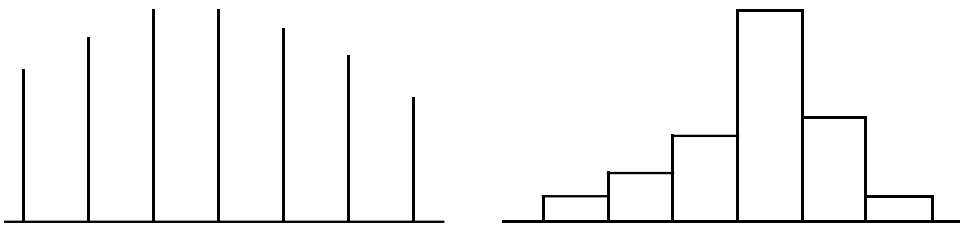
$$g_1 = \frac{m_3}{S^3}$$

erabiltzen da: $g_1 > 0$ bada, maiztasun-banaketa eskuinerantz asimetrikoa dela esaten da; $g_1 < 0$ bada, alderantziz, ezkerrerantz.

II.2.3.2. Kurtosi koefizientea edo zapaltasun/zorroztasun-koefizientea

Simetriarekin batera, itxuraren beste balio tipiko bat ikusiko dugu, kurtosi koefizientea hain zuzen.

Maiztasun-banaketaren grafikoa kontutan hartuz, konkretuki, maiztasun--poligonoa ezaugarri jarraiaren kasuan, edo barra-diagramaren goiko puntuak lotuko lituzkeen lerroa ezaugarri diskretuaren kasuan, kasu hauek dauzkagu:



zapala zorrotza

Batezbesteko aritmetikoa eta bariantza berdina daukaten bi banaketa aztertuz, itxuraren aldetik bata oso zapala da eta oso zorrotza bestea.

Banaketaren maiztasunak batezbesteko aritmetikoaren inguruan handiak baldin badira, momentu zentratu bikoitiek, (adibidez, laugarren ordenakoa), balio handiagoak hartuko dituzte.

Orduan, m_4 zapaltasunaren adierazle bezala har daiteke, baina neurri--unitateak ber lau izanez, zuzenkiago g_2 ikurraz adierazten den kurtosi koefizientea, zapaltasunaren adierazle bezala erabiliko dugu.

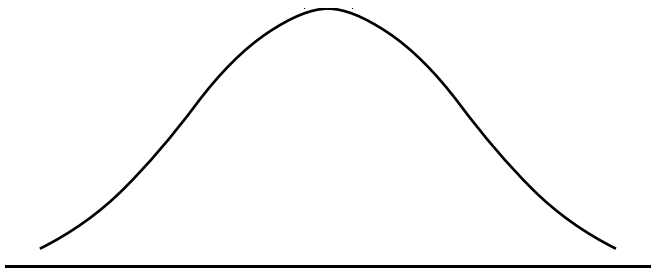
"Normal" deitutako banaketaren kurtosi koefizientea 3 da eta balio hau kurtosia aztertzean konparazio-puntu bezala hartuko dugu.

$$g_2 = \frac{m_4}{S_x^4} = \frac{\sum_{i=1}^{m} (x_i - x)^4 n_i}{N \; S_x^4}$$

g_2, 3 baino handiagoa baldin bada, banaketa zorrotza izango da eta 3 baino txikiagoa baldin bada, zapala izango da.

g_2–3, kurtosi koefizientetzat hartzen badugu, kurtosi positiboa edo negatiboa kontsidera daiteke.

Oharra: Aipatu dugun banaketa "Normala" 3. kurtsoan ikasiko da, inferentzia estatistikoan, oinarrizko probabilitate-eredua baita.



banaketa Normala : $g_2 = 3$

II.3. ALDAGAIEN TRANSFORMAZIO LINEALAK

X aldagaia U aldagaiaren transformatu lineala izan daiteke, hau da:

$$X = a\,U + b$$

b = 0 baldin bada, unitate-aldaketa daukagu soilik

a = 1 baldin bada, jatorri-aldaketa edo traslazioa daukagu soilik, zeina berdin b baita.

Batzutan, jatorria eta unitatea aldatzea komenigarria da, honela, kalkuluak errazten baitira.

Ikus ditzagun, bi aldagaien estatistiko garrantzitsuenen arteko erlazioak:

$$\bar{x} = \frac{1}{N}\sum_i x_i n_i = \frac{1}{N}\sum_i (au_i + b)n_i =$$

$$= a \cdot \frac{1}{N}\sum_i u_i n_i + b\frac{1}{n}\sum_i n_i = a\bar{u} + b$$

Ikusten dugunez X eta U aldagaien batezbesteko aritmetikoak ez dira berdinak; pentsa dezagun estatistiko hau, ezaugarriak neurtzen ditugun unitateetan neurri bat dela.

Bariantza eta desbidazio tipikoen arteko erlazioak:

$$S_x^2 = \frac{1}{N}\sum_i \left(x_1 - \bar{x}\right)^2 n_i = \frac{1}{N}\sum_i \left(au_i + b - a\bar{u} - b\right)^2 n_i =$$

$$= a^2 \cdot \frac{1}{N}\sum_i \left(u_i - \bar{u}\right)^2 n_i = a^2 \cdot S_u^2$$

eta $S_x = |a|S_u$

Eta orokorki, momentu zentratuen arteko erlazioak:

$$m_h(x) = \frac{1}{N}\sum_i (x_i - \bar{x})^h n_i = \frac{1}{N}\sum_i (au_i + b - a\bar{u} - b)^h n_i =$$

$$= a^h \frac{1}{N}\sum_i (u_i - \bar{u})^h n_i = a^h m_h(u)$$

Ikusten dugunez, unitate-aldaketak eragina dauka bariantza eta edozein momentu zentratuan baina traslazioak ez.

II.3.1. Aldagai zentratua

Aldagai baten batezbesteko aritmetikoa zero baldin bada, **aldagai zentratua** deitzen da; aldiz, batezbesteko aritmetikoa zero eta desbidazio tipikoa bat balioak baldin badira, **aldagai tipifikatua** deritzogu.

Edozein X aldagairen kasuan bere batezbestekoa kenduz $Z = X - \bar{x}$, aldagai zentratua lortzen dugu.

$$\bar{z} = \frac{1}{N} \sum_i z_j n_i = \frac{1}{N} \sum_i (x_i - \bar{x}) n_i =$$

$$= \frac{1}{N} \sum_i x_j n_i - \frac{1}{N} \bar{x} \sum_i n_i = \bar{x} - \bar{x} = 0$$

Kontutan hartu behar dugu zentratze-eragiketak ez duela unitate-aldaketarik suposatzen, jatorri-aldaketa edota aldagaiaren translazioa ardatzean suposatzen du bakarrik.

Aldagaiaren grabitate-zentrua, translazio honetan, jatorrira eraman dugu, $\tilde{z} = 0$ baita.

II.3.2. Aldagai tipifikatua edo standardizatua

Modu berean, edozein X aldagairen kasuan batezbestekoa kenduz eta desbidazio tipikoaz zatituz $T = (X - \bar{x})/ S_x$, aldagai tipifikatua lortzen dugu.

$$\bar{t} = \frac{1}{S_x} \sum_i (x_i - \bar{x}) \frac{n_i}{N} = 0$$

$$S_t^2 = \frac{1}{N} \sum_i (t_i - \bar{t})^2 n_i = \frac{1}{N} \sum_i \frac{(\bar{x}_i - \bar{x})^2}{S_x^2} n_i =$$

$$= \frac{1}{S_x^2} \frac{1}{N} \sum_i (\bar{x}_i - \bar{x})^2 n_i = \frac{1}{S_x^2} S_x^2 = 1$$

Tipifikazio-eragiketak, zentratzeaz gainera unitatez aldatzea suposatzen du. Aldagai bakoitzaren balio zentratuak, bere desbidazio tipikoaz zatitzean, honen arabera neurturik ditugu, jatorrizko neurri-unitateak desagertu egiten direlarik.

II.4. KONTZENTRAZIO-NEURKETAK

Batezbesteko aritmetikoa egitean zenbakitzailean ateratzen zaigun kopuruak, abiapuntu estatistiko batetatik ikusita, batzutan ez du zentzu argirik eta interesgarririk. Adibidez, pertsona batzuren altuera-banaketa aztertzean, zenbakitzaile hori altueren batuketa izango da; baina, beste kasu batzutan, bereziki

ezaugarri sozioekonomikoetan, aztergarria izaten da, adibidez, alokairu-banaketa aztertzean, alokairu guztien kopurua edo "alokairu-masa" da.

Kontzentrazio-neurketek, zehazki, "masa" baten uniformetasuna neurtzea dute helburutzat.

Normalki errenta eta alokairu-banaketetan erabili ohi dira, baina beste edozein aldagairen banaketatan erabil daitezke.

Suposa dezagun alokairu-banaketa batetan langile guztiek alokairu berdina jasotzen dutela; orduan, banatzearen uniformetasuna osoa litzateke. Baina, ordez, alokairu-masa guztia langile batek jasoko balu, **uniformetasunik eza** ere osoa izango litzateke edo **kontzentrazioa maximoa** dela esango genuke kasu honetan.

Banaketak uniformetasun absoluturantz jotzen duenean, maiztasun-–banaketaren batezbesteko aritmetikoa guztiz adierazgarria da, eta, alderantziz gertatuko da, kontzentrazioa maximoa denean.

II.4.1. LORENZ-en kurba

Grafikoki banaketen kontzentrazioa ikusarazten digun neurrian, interesgarria zaigu.

Lorenz-en kurba eraikitzeko, ondoko taulan ikusten ditugun zutabeak kalkulatu beharko ditugu.

x_i	n_i	$x_i\, n_i$	N_i	M_i	P_i	Q_i
x_1	n_1	$x_1\, n_1$	N_1	M_1	P_1	Q_1
x_2	n_2	$x_2\, n_2$	N_2	M_2	P_2	Q_2
.	.	.	.	.	.	.
.	.	.	.	.	.	.
x_i	n_i	$x_i\, n_i$	N_i	M_i	P_i	Q_i
.	.	.	.	.	.	.
.	.	.	.	.	.	.
x_k	n_k	$x_k\, n_k$	N	M	100	100
	$\sum_i n_i = N$	$\sum_i x_i n_i = M$				

Azken zutabeak honela kalkulatzen dira:

N_i: (n_i) zutabea metatuz
hots: $N_i = n_1 + n_2 \ldots\ldots + n_i$

M_i: $(x_i n_i)$ zutabea metatuz
hots: $M_i = x_1 n_1 + x_2 n_2 + \ldots\ldots + x_i n_i$

$P_i = \frac{N_i}{N}$ 100 maiztasunen (langileena esate baterako) portzentaia metatua

$Q_i = \frac{M_i}{M}$ 100 masaren portzentaia metatua.

$\{(P_i, Q_i)\}_{i \in \mathbb{I}}$ bikote guztiak sistema cartesiar batetan grafikoki adieraziz, eta $(P_1\ Q_1)$, (0,0) jatorriarekin eta beste bikote guztiak elkarren artean zuzenen bitartez lotuz LORENZ-en kurba lortzen dugu.

Adibidez: Suposa dezagun enpresa batetako langileen banaketa bere asteroko alokairuaren arabera (milaka pezetatan) ondoko taularen lehenengo bi zutabeetan emana dela:

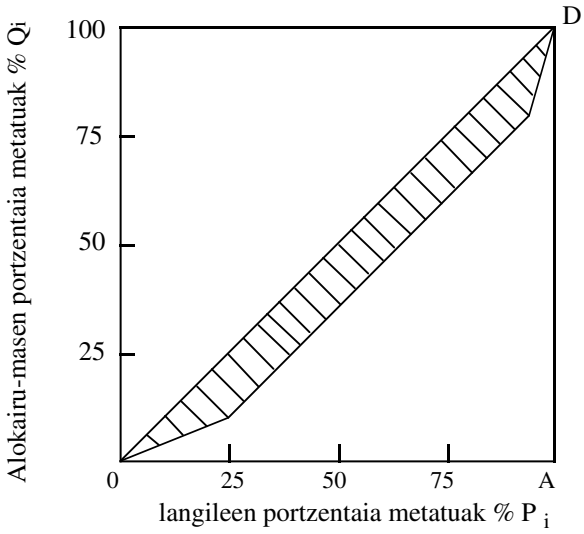
Alokairuak: x_i	Langile kopurua: n_i	Alokairu-masa: $x_i n_i$	Metaketak N_i	Metaketak M_i	Portzentaia metatuak: P_i	Portzentaia metatuak: Q_i
8	10	80	10	80	20	10
15	20	300	30	380	60	47,5
20	15	300	45	680	90	85
24	5	120	50	800	100	100

$N = 50 \qquad M = \sum_{i=1}^{k} x_i n_i = 800$

Adibideari dagokion LORENZ-en kurba ondoko hau izango da.

Eskalak, bi ardatzetan berdinak direnez gero, LORENZ-en kurba, karratu baten barruan aurkitzen da beti. Jatorritik irteten den OD bere diagonala erabiliko dugu uniformetasunaren erreferentzia moduan ($P_i\ Q_i$) multzoko puntu guztiak, OD-n aurkitzen badira, orduan uniformetasun osoa baitugu.

Hots: langileen arteko % 10-ak, alokairu-masaren % 10-a jasotzen du, % 30-ak, alokairu-masaren % 30-a, e.a.



Zenbat eta tarte handiagoa egon kurba eta diagonalaren artean, orduan eta **banaketa kontzentratuagoa** izango da, (edota zenbat eta kurba diagonaletik hurbilago egon, orduan eta **banaketa uniformeagoa**).

Abzisa ardatzean P_i kokatzen badugu eta ordenatu ardatzean Q_i , LORENZ--en kurba, karratuaren beheko triangeluan edukiko dugu (alokairuak txikienetatik handienetarantz ordenatu ditugunez gero, adibidez langileen % 25-ak ezin du alokairu-masaren % 25-a inoiz lortu).

II.4.2. GINI-ren indizea

LORENZ-en kurba banaketen kontzentrazioaren argitzaile bada ere, zenbakizko indize batetaz kontzentrazioa neurtzea, komenigarria zaigu; esate baterako, erabilgarria izango zaigu banaketak konparatzerakoan.

Helburu honi erantzuten dio **Gini-ren indizeak** edo **kontzentrazio-indizeak**.

Marraztutako azalera diagonal eta kurbaren artean dagoena bada, honela definituko dugu I_G, GINI-ren indizea:

$$I_G = \frac{\text{Marraztutako azalera}}{\overset{\triangle}{\text{OAD}} \text{ azalera}}$$

non $0 \leq I_G \leq I$

Kontzentrazioa maximoa bada, orduan $\overset{\triangle}{OAD}$ azalera = marraztutako azalera, eta $I_G = 1$.

Ordez, kontzentrazioa minimoa bada edota banaketa uniformea, orduan marraztutako azalera = 0 izango da eta $I_G = 0$.

Kalkulua: I_G GINI-ren indizea kalkulatzean, azalera horiek zuzenki neurtuz kalkula litezke; bainan beti era honetan erraza izaten ez denez gero, **gutxi gora-behera** ondoko formula erabiliz kalkulatzen da I_G :

$$I_G = \frac{\sum_{i=1}^{k-1} (P_i - Q_i)}{\sum_{i=1}^{k-1} P_i}$$

Kontura gaitezen, batukariak k-1 batugai bakarrik dituela, $p_k - q_k = 0$ baita.

Gutxi gora-beherako formula honetaz ere egiazta daiteke $0 \leq I_G \leq 1$ dela.

Adibidez: Lehenengo taularen datuekin jarraitzen badugu, aurreko formula erabiliz, I_G GINI-ren kontzentrazio-indizea ondoko hau izango da:

P_i	Q_i	$P_i - Q_i$
20	10	10
60	47'5	12'5
90	85	5
170		27'5

$$I_G = \frac{\sum_{i=1}^{k-1} (P_i - Q_i)}{\sum_{i=1}^{k-1} P_i} = \frac{27'5}{170} = 0'16$$

LORENZ-en kurba nahiz GINI-ren indizea aldagai sozioekonomikoetarako erabilgarriak dira. Adibide bezala hauek aipatuko ditugu: errenta pertsonalaren banaketa, nekazal-propietatearen banaketa, sektore batetako enpresa--produkzioaren banaketa eta abar.

III. EZAUGARRI ESTATISTIKO BIKOITZAK, BANAKETAK, TAULAK, ADIERAZPIDE GRAFIKOAK

III.1. EZAUGARRI ESTATISTIKO BIKOITZAK

III.2. MAIZTASUN-BANAKETA BIKOITZAK

III.2.1. Maiztasun-banaketa bikoitzak eta bazter-maiztasunak

III.2.2. Maiztasun-banaketa bikoitzak: Taulak

III.3. MAIZTASUN-BANAKETA BIKOITZEN ADIERAZPIDE GRAFIKOAK

III.3.1. Sakabanatze-diagrama edo puntu-hodeia

III.4. MAIZTASUN-BANAKETA BALDINTZATUAK

III.5. TAULA BATEN DEPENDENTZIA EDO INDEPENDENTZIA

III.6. KONTINGENTZI TAULA BATEN ERRENKADA ETA ZUTABEEN BATEZBESTEKO SOSLAIAK

III.1. EZAUGARRI ESTATISTIKO BIKOITZAK

Orain arte ikusi ditugun fenomenoetan, indibiduo bakoitzari ezaugarri estatistiko bakoitzarekiko neurketa bat besterik ez zitzaion egiten.

Orain aipatuko ditugun ikerketetan, lagineko **indibiduo bakoitzari bi neurketa** egiten dizkiogu, bakoitza ezaugarri estatistiko bati dagokiona.

Ezaugarri estatistiko bikoitzak edo ezaugarri estatistiko bidimentsionalak ikusiko ditugu, bada, ondoren.

Adibidez: Urte batetan soldaduzkara doazen gazteei bi neurketa egiten zaizkie, edo beraien bi ezaugarri aztertzen dira batera: pisua eta altuera.

III.2. MAIZTASUN-BANAKETA BIKOITZAK

III.2.1. Maiztasun-banaketa bikoitzak eta bazter-maiztasunak

Suposa dezagun N indibiduoz osatutako lagin bat, non X ezaugarri batek $x_1\ x_2\ .....\ x_i\ .....\ x_k$ balio desberdinak hartzen dituen.

Suposa dezagun, halaber, beste Y ezaugarri bakun batek lagin berdineko indibiduoetan $y_1\ y_2\ .....\ y_j\ .....\ y_l$ balio desberdinak hartzen dituela.

Orduan, (X,Y) ezaugarri bikoitzak, lagineko indibiduo bakoitzean balio--bikote bat hartzen du.

Hots:

$\omega_p \rightarrow (x_i, y_j)$ $\quad p = 1,....., N$ $\quad i = 1,....., k$ $\quad j = 1,....., l$

Eta (X,Y) ezaugarri bikoitzak, laginean hartuko lituzkeen **balio-bikote desberdinen** kopurua, k x l izango da. Hemendik aurrera horiek bakarrik hartuko ditugu kontuan.

Hots:

$$\{\omega_1, \omega_2 ..... \omega_N\} \xrightarrow{(X,Y)} \{(x_1 y_1)\ (x_1 y_2) ..... (x_i y_1) ..... \ ..... \ (x_i y_l) ..... (x_k y_1) ..... \ ..... (x_k y_l)\} = \{(x_i y_j)\}$$

(x_i y_j) **balio-bikotearen MAIZTASUN ABSOLUTUA**

Sinbolikoki n_{ij} ikurraz adieraziko dugu, eta laginean (x_i y_j) balio-bikotea hartzen duten indibiduoen kopurua da.

Non:

$$\sum_{i=1}^{l} \sum_{i=1}^{k} n_{ij} = n_{..} = N$$

(x_i y_j) **balio-bikotearen MAIZTASUN ERLATIBOA**

Sinbolikoki f_{ij} ikurraz adieraziko dugu, non:

$$f_{ij} = \frac{n_{ij}}{N} \qquad \sum_{j=1}^{l} \sum_{i=1}^{k} f_{ij} = 1$$

eta portzentaietan:

$$p_{ij} = \%f_{ij} \cdot 100 \qquad \sum_{j} \sum_{i} p_{ij} = 100$$

x_i **balioaren BAZTER-MAIZTASUN ABSOLUTUA**

Sinbolikoki $n_{i.}$ ikurraz adieraziko dugu, non:

$$n_{i.} = \sum_{j=1}^{l} n_{ij} \qquad \sum_{j} n_{i.} = n_{..} = N$$

y_j **balioaren BAZTER-MAIZTASUN ABSOLUTUA**

Sinbolikoki $n_{.j}$ ikurraz adieraziko dugu, non:

$$n_{.j} = \sum_{j=1}^{k} n_{ij} \qquad \sum_{j} n_{.j} = n_{..} = N$$

x_i **balioaren BAZTER-MAIZTASUN ERLATIBOA**

Sinbolikoki $f_{i.}$ ikurraz adieraziko dugu, non:

$$f_{.i} = \frac{n_{i.}}{N} = \frac{\sum_{j} n_{ij}}{N}$$

y_j **balioaren BAZTER-MAIZTASUN ERLATIBOA**

Sinbolikoki $f_{.j}$ ikurraz adieraziko dugu, non:

$$f_{.j} = \frac{n_{.j}}{N} = \frac{\sum_j n_{ij}}{N}$$

eta

$$f_{..} = \sum_j f_{i.} = \sum_j f_{.j} = 1$$

III.2.2. Maiztasun-banaketa bikoitzak: Taulak

Lagin bati dagozkion maiztasun-banaketen taulak, ezaugarria bikoitza bada, ondoan azaltzen ditugunak dira.

X \ Y	y_1	y_2	..	..	y_j	..	..	y_l	
x_1	n_{11} f_{11}	n_{12} f_{12}	..	..	n_{1j} f_{1j}	..	..	n_{1l} f_{1l}	$n_{1.}$ $f_{1.}$
x_2	n_{21} f_{21}	n_{22} f_{22}	..	..	n_{2j} f_{2j}	..	..	n_{2l} f_{2l}	$n_{2.}$ $f_{2.}$
. .			.. ..	.. ..		.. ..	.. ..		
x_i	n_{i1} f_{i1}	n_{i2} f_{i2}	.. ..	.. ..	n_{ij} f_{ij}	.. ..	.. ..	n_{il} f_{il}	$n_{i.}$ $f_{i.}$
. .	. .	. .	.. ..	.. ..	. .	.. ..	.. ..	. .	. .
x_k	n_{k1} f_{k1}	n_{k2} f_{k2}	..	..	n_{kj} f_{kj}	..	..	n_{kl} f_{kl}	$n_{k.}$ $f_{k.}$
Guztira	$n_{.1}$ $f_{.1}$	$n_{.2}$ $f_{.2}$			$n_{.j}$ $f_{.j}$			$n_{.l}$ $f_{.l}$	$n_{..}=N$ $f_{..}=1$

BAZTER-BANAKETAK

X	M. Ab.	M. Er.
x_1	$n_{1.}$	$f_{1.}$
x_2	$n_{2.}$	$f_{2.}$
:	:	:
x_i	$n_{i.}$	$f_{i.}$
:	:	:
:	:	:
x_k	$n_{k.}$	$f_{k.}$

Y	M. Ab.	M. Er.
y_1	$n_{.1}$	$f_{.1}$
y_2	$n_{.2}$	$f_{.2}$
:	:	:
y_j	$n_{.j}$	$f_{.j}$
:	:	:
:	:	:
y_l	$n_{.l}$	$f_{.l}$

$$\sum_i n_{i.} = N \qquad \sum_j n_{.j} = N$$

$$\sum_i f_{i.} = 1 \qquad \sum_j f_{.j} = 1$$

(X edo Y ezaugarriak, edo biak, jarraiak badira, ezaugarri bakunetan bezala mailakatuz diskretu bihurtuko ditugu).

Adibidea: 750 familia duen talde batean bi ezaugarri aztertu dira batera. Bata aktibo diren pertsonen kopurua, eta bestea hileroko soldata (milaka pezetatan).

Ondoko taulan dauzkagu maiztasun absolutuen banaketak.

Hileroko soldata milaka pezetatan

Pertsona aktiboak	(60,80)	(80,140)	(140,250)	Guztira
1	143	59	--	202
2	46	170	32	248
3	----	93	86	179
4	----	17	104	121
Guztira	189	339	222	750

Bazter-maiztasunaren esangura, bereziki azpimarratzea komeni zaigu.

Ikus dezagun nola maiztasun absolutuen kopuruetatik (ezaugarri bikoitzari dagozkionak), batuketa batzu eginez, bazter-maiztasunen banaketak lortzen

ditugun: hau da, errenkadak batuz, ezaugarri batenak (pertsona aktiboak), eta zutabeak batuz, beste ezaugarriarenak (hileroko soldata).

Ohargarria da, bada, bazter-maiztasunen banaketak, maiztasun bakunen banaketa batzu besterik ez direla.

III.3. MAIZTASUN-BANAKETA BIKOITZEN ADIERAZPIDE GRAFIKOAK

Aldagai bikoitz baten maiztasun-banaketa ikustarazteko bi adierazpide grafiko ditugu, ESTEREOGRAMA eta SAKABANATZE-DIAGRAMA edo PUNTU-HODEIA hain zuzen. Hauetariko bigarrena ondoan azalduko dugu, bestea ez baita ia erabiltzen.

III.3.1. Sakabanatze-diagrama edo puntu-hodeia

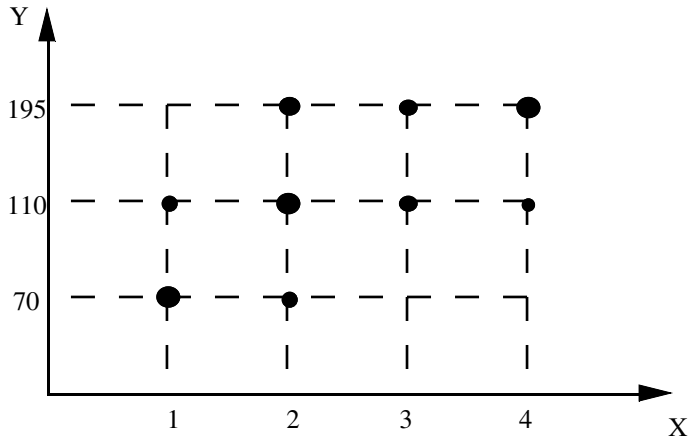
Izan bedi (X,Y) ezaugarri estatistiko bikoitz bat, eta izan bitez

$\left\{\left(x_i y_j\right)\right\}$ $j = 1 \ldots\ldots 1$ (X,Y) ezaugarriak N tamainuko lagin batetan hartzen
$i = 1 \ldots\ldots k$ dituen balio-bikote desberdinak.

Orduan, ardatz cartesiarreko sistema batetan $(x_i y_j)$ balio-bikote bakoitza puntu batez adierazten badugu,

$\left\{\left(x_i y_j\right)\right\}$ $j = 1,\ldots\ldots 1$ multzo osoaren adierazpenari, sakabanatze-diagrama
$i = 1,\ldots\ldots k$ edo puntu-hodeia deritzo.

Ondoan, 2.2. ataleko adibideari dagokion puntu-hodeia daukagu.



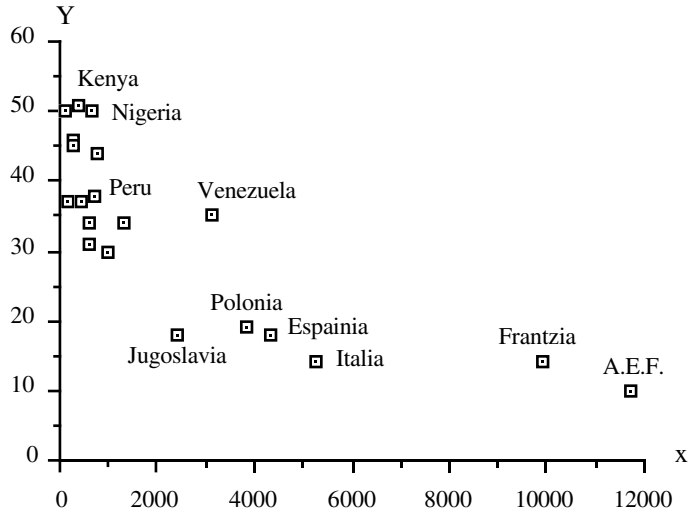
Ikusten dugunez, $(x_i y_j)$ balio-bikoteak errepikatzen direnean sakabanatze--diagramako puntuetariko batzu besteak baino lodiagoak dira, beraien azalerak eta balio-bikoteei dagozkien maiztasunak proportzionalak izanik.

Dena dela, taula bikoitzak edo datu taldekatuen taulak eta beraren dispertsio--diagramak informazio berdintsua ematen digute.

Batez ere datu-multzo handiak ditugunean, daturik isolatuenak edo ezer gutxi errepikatzen diren datuenak dira taulak; eta horrelako kasuetan, puntu-hodeiak oso baliagarriak zaizkigu, multzoaren ibilbidea ikustarazteko.

X ezaugarria Nazio-Produktu Gordina eta Y jaiokortasuna izanik, ondoan 20 estaturi dagozkien datuak eta beraien puntu-hodeia ditugu.

ESTATUA	IZENA	NPG	JAIOKORTASUNA
1	Ethiopia	130	50
2	Birmania	160	37
3	Tanzania	270	46
4	Uganda	290	45
5	Kenya	380	51
6	Egipto	460	37
7	Thailandia	590	31
8	Filipinas	600	34
9	Nigeria	670	50
10	Peru	730	38
11	Maroko	740	44
12	Kolonbia	1010	30
13	Turkia	1330	34
14	Jugoslavia	2430	18
15	Venezuela	3130	35
16	Polonia	3830	19
17	Espainia	4340	18
18	Italia	5240	14
19	Frantzia	9940	14
20	A. Err. Federatua	11730	10



III.4. MAIZTASUN-BANAKETA BALDINTZATUAK

(X,Y) ezaugarri edo aldagaien banaketa bikoitzaren kasuan eta X aldagaiaren x_i balioarentzat, Y aldagaiaren maiztasun-banaketa kontsidera daiteke.

X \ Y	y_1	y_2	.....	y_j	.....	y_l	
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	
x_i	n_{i1}	n_{i2}	.....	n_{ij}	.....	n_{il}	$n_{i.}$
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	

x_i balioa baldintzatzailea izanez, maiztasun-banaketa baldintzatuak ondokoak ditugu:

$Y/X = x_i$	$n_{j/i}$	$f_{j/i}$
y_i	$n_{1/i} = n_{i1}$	$\frac{n_{1/i}}{n_{i.}}$
y_2	$n_{2/i} = n_{i2}$	$\frac{n_{2/i}}{n_{i.}}$
.	.	.
.	. **(1)**	. **(2)**
.	.	.
y_j	$n_{j/i} = n_{ij}$	$\frac{n_{j/i}}{n_{i.}}$
.	.	.
.	.	.
.	.	.
y_l	$n_{l/i} = n_{il}$	$\frac{n_{l/i}}{n_{i.}}$

eta (**1**) x_i balioaz baldintzatutako Y aldagaiaren maiztasun absolutuen banaketa deitzen da.

Maiztasun-banaketa honen balioak, $n_{i.}$ bazter-maiztasunaz zatitzen baditugu (**2**), maiztasun-banaketa erlatiboa lortzen dugu, eta:

$$f_{i/j} = \frac{n_{j/i}}{n_{i.}} = \frac{n_{ij}}{n_{i.}}$$

X aldagaiak hartzen dituen balio guztiak kontsideratuz, beste hainbeste Y aldagaiaren banaketa baldintzatu edukiko ditugu.

Analogikoki, X aldagaiaren banaketa baldintzatuak edukiko ditugu, Y aldagaiak hartzen dituen balio guztiak baldintzatzaileak izanez.

Adibidea:

Ondoko taulan (X,Y) aldagaien banaketa daukagu. Y aldagaiak pieza baten luzera adierazten du.

X aldagaiak, ordea, akastuna (B) edo akasgabea (A) izatea.

X \ Y	71	72	73	$n_{i.}$
B	3	1	5	9
A	2	4	6	12
$n_{.j}$	5	5	11	21

Maiztasun-banaketa baldintzatuak (erlatiboak) hauek dira:

f (y/x = B)	3/9	1/9	5/9
f (y/x = A)	2/12	4/12	6/12

f(x/y = 71)	f(x/y = 72)	f(x/y = 73)
3/5	1/5	5/11
2/5	4/5	6/11

Horrelako banaketa baldintzatuek zera adierazten digute: aldagai baten balio bat finko kontsideratuz, beste aldagaiaren balioek dauzkaten proportzioak zeintzu diren.

Hau da: pieza horiek osatzen duten populazioan akastunen azpipopulazioa kontsideratuz, zer proportziotan daude luzera desberdinetakoak akastunen artean? Erantzuna f(y/x = B) banaketak ematen digu.

Analogikoki, akasgabeen azpipopulazioa kontsideratzen badugu, f(y/x= A) edukiko dugu.

Aldiz, Y aldagaiaren balio bat baldintzatzailea baldin bada, dagokion azpipopulazioa kontsideratuko dugu.

Adibidez, Y aldagaiak luzera milimetrotan adierazten badu eta 73 mm neurtzen dituztenen azpipopulazioa kontsideratuz, zer proportziotan dira akastunak ala akasgabeak pieza horiek? Erantzuna kasu honetan, f(x/y = 73) banaketak ematen digu.

III.5. TAULA BATEN DEPENDENTZIA EDO INDEPENDENTZIA

X eta Y aldagaiak kontsideratuz, aldagai baten aldaketarekin bestearen banaketa aldatzen bada, biak elkarren artean dependenteak dira.

Dakusagun ondoko taula:

X \ Y	71	72	73	$n_{i.}$	$f_{i.}$
B	3	1	5	9	9/21
A	2	4	6	12	12/21
$n_{.j}$	5	5	11	21	
$f_{.j}$	5/21	5/21	11/21		1

X aldatzean, Y aldagaiaren banaketa aldatzen da, hau da:

y_j/x_i	n(y/x = B)	n(y/x = A)	f(y/x = B)	f(y/x = A)
71	3	2	1/3	1/6
72	1	4	1/9	1/3
73	5	6	5/9	1/2

Luzera horiek, bada, ez daude berdin banatuta akastunen artean eta akasgabeen artean, hots, nolabaiteko menpekotasun edo dependentzia ikusten da taula honetan.

Demagun, orain, honako taula hau:

X \ Y	71	72	73	$n_{i.}$	$f_{i.}$
B	3	6	9	18	18/30
A	2	4	6	12	12/30
$n_{.j}$	5	10	15	30 = N = n..	
$f_{.j}$	5/30	10/30	15/30		

X aldatzean, Y aldagaiaren banaketa ez da aldatzen, hau da:

y_j/x_i	n(y/x = B)	n(y/x = A)	f(y/x = B)	f(y/x = A)
71	3	2	3/18	2/12
72	6	4	6/18	4/12
73	9	6	9/18	6/12

Ikusten dugunez, Y aldagaiaren banaketa baldintzatuak berdinak dira, luzerak proportzio berdinetan daude akastunen eta akasgabeen artean, eta Y aldagaiaren bazter-banaketa (5/30, 10/30, 15/30) ere berdina da.

Hau da, x_i balioaz baldintzatutako y_j balioaren maiztasun erlatibo baldintzatua eta y_j balioaren bazter-maiztasun erlatiboa berdinak dira:

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$$

X aldagaiaren banaketa baldintzatuak eta bazter-banaketa kontsideratuz, gauza bera gertatzen da, hots, Y aldagaiaren balioak aldatzean, X aldagaiaren banaketa ez da aldatzen eta X eta Y independenteki banatzen dira.

Aurreko erlazioaren simetrikoa edukiko dugu, hau da:

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{.i}}{N}$$

Bi erlazio horietatik erraz ateratzen den independentziaren baldintza beharrezkoa eta nahikoa ondokoa da:

$$f_{ij} = f_{i.}\ f_{.j} \qquad \forall(i,j)$$

Hots, edozein (x_i y_j) bikoterentzat, bazter-maiztasun erlatiboen biderkadura eta bikote horren maiztasun erlatiboa berdinak dira.

III.6. KONTINGENTZI TAULA BATEN ERRENKADA ETA ZUTABEEN BATEZBESTEKO SOSLAIAK

X \ Y	Y_1 ..... y_j ..... y_l	$n_{i.}$
x_1	n_{11} .... n_{1j} ....n_{1l}	$n_{1.}$
⋮	⋮ ⋮ ⋮	⋮
x_i	n_{i1} .... n_{ij} .... n_{il}	$n_{i.}$
⋮	⋮ ⋮ ⋮	⋮
x_k	n_{k1} .... n_{kj} ... n_{kl}	$n_{k.}$
$n_{.j}$	$n_{.1}$ .... $n_{.j}$ ... $n_{.l}$	N

Y eta X aldagaien maiztasun erlatiboen bazter-banaketak, Y aldagaiaren banaketa baldintzatuen eta X aldagaiaren banaketa baldintzatuen batezbesteko soslaiak dira. Y aldagaiaren banaketa baldintzatuen edo errenkada soslaien batezbestekoa hauxe da:

$$\sum_i \frac{n_{ij}}{n_{i.}} f_{i.} = f_{.j}$$

$\sum_i \frac{n_{ij}}{n_{i.}} f_{i.} = f_{.j}$ balioen batezbesteko ponderatua da, $f_{i.}$ maiztasuna x_i elementu baldintzatzaile bakoitzaren pisu erlatiboa izanik.

Halaber, $\sum_i \frac{n_{ij}}{n_{.j}} f_{.j} = f_{.j}$

Hots: $f_{i.}$, $\frac{n_{ij}}{n_{.j}}$ balioen batezbesteko ponderatua da $f_{.j}$ maiztasuna y_j elementu baldintzatzaile bakoitzaren pisu erlatiboa izanik.

IV. EZAUGARRI BIKOITZEN BALIO TIPIKOAK

IV.1.MOMENTUAK
IV.1.1. Momentu arruntak edo jatorriarekikoak
IV.1.2. Momentu zentralak edo batezbestekoarekikoak
IV.2.MAIZTASUN-BANAKETA BIKOITZEN BALIO TIPIKO EDO ESTATISTIKOAK
IV.2.1. Kobariantza: S_{xy}
IV.2.2. Koerlazio-koefizientea: r_{xy}
IV.3.ALDAGAIEN TRANSFORMAZIO LINEALAK
IV.4.KOERLAZIO GABEKO ALDAGAIEN BI PROPIETATE
IV.5.SAKABANATZE- ETA KOERLAZIO-MATRIZEAK
IV.6.OROKORPENA

IV.1. MOMENTUAK

II. ikasgaian, kasu bakunerako gertatu zen bezala, balio tipikoak edo estatistikoak aurkeztu baino lehen, momentuak ikusi behar ditugu eta arrazoi berdinagatik, hots, balio tipikoak momentuak direlako edo momentuen bidez definitzen direlako.

IV.1.1. Momentu arruntak edo jatorriarekikoak

Maiztasun-banaketa bikoitz baten h. ordenako momentu arrunta edo jatorriarekikoa, ondoko formularen bidez definitzen da:

$$a_{h_1 h_2} = \frac{1}{N} \sum_i \sum_j x_i^{h_1} y_j^{h_2} n_{ij} \qquad h_1 + h_2 = h$$

1. ordenako edo lehenengo ordenako momentu arruntak hauexek dira:

$$a_{10} = \frac{1}{N} \sum_i \sum_j x_i n_{ij} = \frac{1}{N} \sum_i x_i \sum_j n_{ij} =$$

$$= \frac{1}{N} \sum_i x_i n_{i.} = \overline{x}$$

$$a_{01} = \frac{1}{N} \sum_i \sum_i y_j n_{ij} = \frac{1}{N} \sum_j y_j \sum_i n_{ij} =$$

$$= \frac{1}{N} \sum_j y_j n_{.j} = \overline{y}$$

2. ordenako edo bigarren ordenako momentu arruntak hauexek dira:

$$a_{20} = \frac{1}{N} \sum_i \sum_j x_i^2 n_{ij} = \frac{1}{N} \sum_i x_i^2 \sum_j n_{ij} =$$

$$= \frac{1}{N} \sum_i x_i^2 n_{i.} = a_2(x)$$

Hau da, X ezaugarriaren bazter-banaketa kontsideratuz, banaketa bakun honen bigarren ordenako momentu arrunta, aurrekoa da.

$$a_{02} = \frac{1}{N}\sum_i \sum_i y_i^2 n_{ij} = \frac{1}{N}\sum_j y_i^2 \sum_i n_{ij} =$$

$$= \frac{1}{N}\sum_j y_i^2 n_{.j} = a_2(y)$$

Y ezaugarriaren bazter-banaketaren bigarren ordenako momentu arrunta, aurrekoa izango da.

$$a_{11} = \frac{1}{N}\sum_i \sum_j x_i y_j n_{ij}$$

Hirugarren ordenako momentuak hauexek izango lirateke:

$$a_{30},\ a_{03},\ a_{21},\ a_{12}$$

IV.1.2. Momentu zentralak edo batezbestekoarekikoak

Maiztasun-banaketa bikoitz baten h-ordenako momentu zentrala edo batezbestekoarekikoa, ondoko formularen bidez definitzen da:

$$m_{h_1 h_2} = \frac{1}{N}\sum_i \sum_j (x_i - \bar{x})^{h_1} (y_j - \bar{y})^{h_2} n_{ij}$$

Bereziki, bigarren ordenakoak oso garrantzitsuak dira:

$$m_{20},\ m_{02},\ m_{11}$$

$$m_{20} = \frac{1}{N}\sum_i \sum_j (x_i - \bar{x})^2 n_{ij} = \frac{1}{N}\sum_j (x_i - \bar{x})^2 \sum_j n_{ij} =$$

$$= \frac{1}{N}\sum_i (x_i - \bar{x})^2 n_{i.} = S_x^2$$

Hori da, bestetik, X ezaugarriaren bariantza (ikus 37. orrialdea)

$$m_{20} = \frac{1}{N}\sum_i \sum_j (y_i - \bar{y})^2 n_{ij} = \frac{1}{N}\sum_j (y_i - \bar{y})^2 \sum_i n_{ij} =$$

$$= \frac{1}{N}\sum_j (y_j - \bar{y})^2 n_{.j} = S_y^2$$

Y ezaugarriaren bariantza da.

$$m_{11} = \frac{1}{N}\sum_i \sum_j (x_i - \bar{x})(y_i - \bar{y}) n_{ij}$$

m_{11} momentua S_{xy} KOBARIANTZA deitzen dena da, balio tipiko garrantzitsuenetariko bat, hain zuzen.

IV.2. MAIZTASUN-BANAKETA BIKOITZEN BALIO TIPIKO EDO ESTATISTIKOAK

(X,Y) ezaugarri estatistiko bikoitz bat bada, eta bere maiztasun-banaketa lagin batetan kalkulatuta baldin badugu, X eta Y-ren arteko erlazioa neurtzen duten estatistiko batzu kalkula genitzake.

IV.2.1. Kobariantza: S_{xy}

X eta Y ezaugarriek, laginean duten erlazio-neurri bat da (non X eta Y ezaugarriak indibiduoen gain neurtzen direneko unitateek eragina duten).

Esan dugunez, m_{11} bigarren ordenako momentu hau kobariantza da; ikus dezagun nola adierazten den momentu arrunten bidez:

$$S_{xy} = m_{11} = \frac{1}{N}\sum_i \sum_j (x_i - \bar{x})(y_i - \bar{y}) n_{ij} =$$

$$= \frac{1}{N}\sum_i \sum_j (x_i y_j - x_i \bar{y} - y_j \bar{x} + \bar{x}\bar{y}) n_{ij} =$$

$$= \frac{1}{N}\sum_i \sum_j x_i y_j n_{ij} - \frac{1}{N}\bar{y} \sum_i \sum_j x_i n_{ij} -$$

$$-\frac{1}{N}\bar{x}\sum_i\sum_j y_j n_{ij} + \frac{1}{N}\bar{x}\bar{y}\sum_i\sum_j n_{ij} =$$

$$= a_{11} - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} = a_{11} - \bar{x}\bar{y}$$

IV.2.2. Koerlazio-koefizientea: r_{xy}

Estatistiko honek ere X eta Y aldagaiek laginean duten erlazioa neurtzen digu, bainan kasu honetan berdin da X eta Y aldagaien balioak zein unitatetan neurtzen diren, neurketan erabiltzen diren unitateek ez baitute eraginik koerlazio-koefizientean.

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_i\sum_j (x_i - \bar{x})(y_j - \bar{y}) n_{ij} / N}{\left(\frac{1}{N}\sum_i (x_i - \bar{x})^2 n_{i.}\right)^{1/2}\left(\frac{1}{N}\sum_j (y_j - \bar{y})^2 n_{.j}\right)^{1/2}} =$$

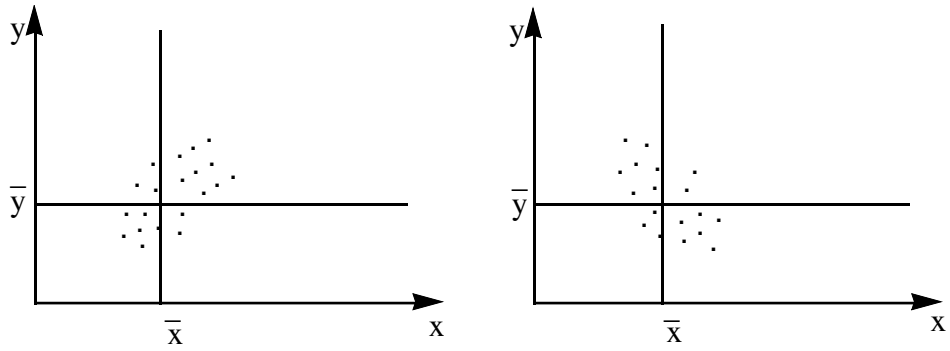
$$= \frac{a_{11} - \bar{x}\bar{y}}{(a_{20} - \bar{x}^2)^{1/2}(a_{02} - \bar{y}^2)^{1/2}}$$

Dakigunez S_x eta S_y desbidazio tipikoak, beti positiboak dira; orduan, S_{xy} kobariantza positiboa (negatiboa) baldin bada, r_{xy} koerlazio-koefizientea positiboa (negatiboa) izango da.

Intuitiboki honetaz ohar daiteke: puntu gehienentzako $(x_i - \bar{x})$ eta $(y_j - \bar{y})$ kendurak biak positiboak edo negatiboak baldin badira, (hau da, $(\bar{x}, \bar{y})$ puntua jatorritzat hartzen badugu, sakabanatze-diagramako puntuak 1. eta 3. koadranteetan egongo dira) $S_{xy} > 0$ izango da; aldiz, puntu gehienentzako zeinu kontrakoak baldin badira (puntuak 2. eta 4. koadranteetaan egongo dira), $S_{xy} < 0$.

Hots: kobariantzaren zeinuak puntu-banaketaren ideia bat ematen digu.

Grafikoki:



Kobariantzak nahiz koerlazio-koefizienteak bi ezaugarrien artean dagoen erlazioa adierazten digute; baina ikusi dugu zein kasutan izango den $S_{xy} > 0$, hots, $(x_i - \bar{x})$ eta $(y_j - \bar{y})$ kendurak biak positiboak direnean (hau da, bi ezaugarriak gorakorrak dira) edo biak negatiboak direnean (hau da, bi ezaugarriak beherakorrak). $S_{xy} < 0$ den kasuan, kenduren zeinuak kontrakoak direnez, ezaugarri bat gorakorra da eta beherakorra bestea. Azkenez $S_{xy} = 0$ baldin bada, bi ezaugarriek ez dute erlazio linealik elkarren artean, hau da, koerlazio gabeak dira.

Ohar daiteke taula baten independentziak koerlazio eza halabehartzen duela, baina ez alderantziz.

$$\text{Independentzia } (f_{ij} = f_{i.}\ f_{.j}) \Rightarrow \not\Leftarrow \text{ koerlazio eza}$$

Hots: Independentziaren kasuan

$$S_{xy} = \sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) f_{i.} f_{.j} = \sum_i (x_i - \bar{x}) f_{i.}$$

$$\sum (y_j - \bar{y}) f_{.j} = m_1(x) \cdot m_1(y) = 0$$

r_{xy} koefizienteak ez du bakarrik aldagaien arteko erlazioa nolakoa den esaten (r_{xy} -ren zeinua S_{xy} -rena baita) baizik eta zer mailatan erlazionatuta dauden ere adierazten du.

Koerlazio-koefizienteak har ditzakeen balioak, [-1, + 1] tartekoak dira.

Zerora hurbiltzen den balio bat hartzen badu, bi ezaugarrien arteko erlazioa ahula da.

r_{xy} koefizientea +1 baliora hurbiltzen bada, erlazioa sakona eta zuzena da.

Aldiz, -1 baliora hurbiltzen bada, erlazioa sakona da, baina alderantzizkoa.

IV.3. ALDAGAIEN TRANSFORMAZIO LINEALAK

Ikus dezagun, bada, (X,Y) aldagaien ordez, haiekin erlazionatuta dauden (U,V) aldagaiak erabiltzen ditugunean, kobariantza aldatu egiten dela eta koerlazio-koefizientea ez.

Horrexegatik eta baita erregresio-teorian betetzen duen funtzioagatik ere, koerlazio-koefizientea, balio tipikorik garrantzitsuena da bi edo ezaugarri gehiago aztertzen ditugun ikerketetan.

Hots: $x_i = au_i + b$ eta $y_j = ev_j + f$ egiten badugu batezbesteko aritmetikoen arteko erlazioak, honako hauek dira:

$$\bar{x} = a\bar{u} + b \qquad \text{eta} \qquad \bar{y} = e\bar{v} + f$$

Ikus dezagun, koerlazio-koefizientea kalkulatzean erabiltzen dugun desbidazio tipikoetan eta kobariantzan, zer eragina duen aldaketa honek.

$$S_x = |a|S_u \qquad \text{eta} \qquad S_y = |e|S_v$$

$$S_{xy} = \frac{1}{N}\sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) =$$

$$= \frac{1}{N}\sum_i \sum_j (au_i + b - a\bar{u} - b)(ev_j + f - e\bar{v} - f)n_{ij} =$$

$$= a \cdot e \cdot \frac{1}{N}\sum_i \sum_j (u_i - \bar{u})(v_j - \bar{v})n_{ij} = a \cdot e \cdot S_{uv}$$

Jatorria aldatzeak eraginik ez duela ikusten da baina, bai baduela unitate--aldaketa horrek.

Aldiz, koerlazio-koefizientean ez du eraginik ez jatorri aldaketak, ezta unitate-aldaketak ere.

Hots:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{a \cdot e \cdot S_{uv}}{|a| S_u |e| S_v} = \frac{S_{uv}}{S_u S_v} = |r_{uv}|$$

Ikusten dugunez, bi aldagaien arteko koerlazio-koefizientea inbariantea da edozein transformazio linealen aurrean.

a edo e negatiboa balitz erlazioaren zentzua aldatuko litzateke.

KOERLAZIO-KOEFIZIENTEA, aldagai tipifikatuen arteko kobariantza da.

Hots: $Z = \frac{(X - \bar{x})}{S_x}$ eta $Z' = \frac{(Y - \bar{y})}{S_y}$ baldin badira:

$$S_{zz'} = \frac{1}{N} \sum_i \sum_j \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_j - \bar{y}}{S_y} \right) n_{ij} =$$

$$= \frac{1}{S_x S_y} S_{xy} = r_{xy}$$

Adibidea: Lagin batetan X eta Y ezaugarriak aztertu dira. Maiztasun--banaketaren taula ondokoa bada, kalkula ditzagun: batezbestekoak, bariantzak, kobariantza eta koerlazio-koefizientea, kalkulurik laburrena eginez.

X \ Y	1'735	1'740	1'745	$n_{i.}$
239	7	9	9	25
243	8	7	5	20
247	11	9	10	30
251	4	7	14	25
$n_{.j}$	30	32	38	100

Aldaketa hauek egiten ditugu: $u_i = \dfrac{x_i - 243}{4}$, $v_j = \dfrac{y_j - 1'740}{0'005}$

hots: $x_i = 4u_i + 243$ eta $y_j = 0'005\, v_j + 1'740$

honela beste taula honetara iristen gara:

u \ v	-1	0	1	$n_{i.}$
-1	7	9	9	25
0	8	7	5	20
1	11	9	10	30
2	4	7	14	25
$n_{.j}$	30	32	38	100

Eta kalkuluak eginez:

$$\overline{u} = \frac{\sum_i u_i n_{i.}}{N} = 0'55$$

$$\overline{x} = 4\ \overline{u} + 243 = 245'2$$

$$\overline{v} = \frac{\sum_i v_j n_{.j}}{N} = 0'08$$

$$\overline{y} = 0'005\ \overline{v} + 1'740 = 1'7404$$

$$S_u^2 = 1'2475$$

$$S_x^2 = 4^2 \quad S_u^2 = 19'96$$

$$S_v^2 = 0'6736$$

$$S_y^2 = 0'005^2 \cdot \quad S_v^2 = 1'684 \cdot 10^{-5}$$

$$S_{uv} = 0'126$$

$$S_{xy} = 4 \cdot \ 0'005 \cdot \ S_{uv} = 0'00252$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{S_{uv}}{S_u S_v} = 0'1374$$

Ikusten dugunez, X eta Y ezaugarrien arteko erlazioa zuzena da eta oso ahula; ez dute ia zerikusirik batak eta besteak.

IV.4. KOERLAZIO GABEKO ALDAGAIEN BI PROPIETATE

2.2. atalean esaten genuen bezala, $S_{xy} = 0$ denean ez dago inongo erlazio linealik X,Y aldagaien artean, horrexegatik $r_{xy} = 0$, hau da, koerlazio-koefiziente lineala zero izatean koerlazio gabeak direla esango dugu.

Dakusagun orain bada, koerlazio gabeko aldagaien bi propietateak.

1. X eta Y aldagaiak koerlazio gabeak direnean, beraien baturaren bariantza bariantzen batura da.

$$S_{xy} = 0 \qquad S^2_{(x+y)} = S^2_x + S^2_y$$

Frogapena:

$$S^2_{(x+y)} = \frac{1}{N}\sum_i \left(x_i + y_i - \bar{x} - \bar{y}\right)^2 = \frac{1}{N}\sum_i \left[\left(x_i - \bar{x}\right) + \left(y_i - \bar{y}\right)\right]^2 =$$

$$= \frac{1}{N}\sum_i \left[\left(x_i - \bar{x}\right)^2 + \left(y_i - \bar{y}\right)^2 + 2\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right] =$$

$$= \frac{1}{N}\sum_i \left(x_i - \bar{x}\right)^2 + \frac{1}{N}\sum_i \left(y_i - \bar{y}\right)^2 + 2\frac{1}{N}\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) =$$

$$= S^2_x + S^2_y + 2S_{xy}$$

Hots:

$$S_{xy} = 0 \quad \Rightarrow \ S^2_{(x+y)} = S^2_x + S^2_y$$

2. X aldagaia Y,Z.... aldagaiekin koerlazio gabea denean, beraien edozein konbinazio linealekin koerlazio gabea da.

$$\left.\begin{array}{l} S_{x,y}=0 \\ S_{x,z}=0 \end{array}\right\} \Rightarrow S_{x,ay+bz+c}=0$$

Frogapena:

$$S_{x,ay+bz+c}=\frac{1}{N}\sum_i\left[(x_i-\bar{x})(ay_i+bz_i+c-a\bar{y}-b\bar{z}-c)\right]=$$

$$=\frac{1}{N}\sum_i(x_i-\bar{x})\left[a(y_i-\bar{y})+b(z_i-\bar{z})\right]=a\frac{1}{N}\sum_i(x_i-\bar{x})(y_i-\bar{y})+$$

$$+b\frac{1}{N}\sum_i(x_i-\bar{x})(z_i-\bar{z})=a\,S_{x,y}+b\,S_{x,z}=0$$

IV.5. SAKABANATZE- ETA KOERLAZIO-MATRIZEAK

Momentu zentralak honela ere adierazten dira:

$$l_{11}=m_{20}=S_x^2,\ l_{22}=m_{02}=S_y^2,\ l_{12}=l_{21}=m_{11}=S_{xy}$$

L eta **R** letraz laburki adierazten ditugun ondoko matrizeak, sakabanatze-edo kobariantza matrizea eta koerlazio-matrizeak dira.

$$\mathbf{L}=\begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix} \qquad \mathbf{R}=\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}$$

Ikusten denez, simetrikoak dira biak.

Lehengo adibidekoak hauek dira:

$$\mathbf{L}=\begin{bmatrix} 19'96 & 0'0025 \\ 0'0025 & 1'684\cdot 10^{-5} \end{bmatrix} \qquad \mathbf{R}=\begin{bmatrix} 1 & 0'1374 \\ 0'1374 & 1 \end{bmatrix}$$

IV.6. OROKORPENA

Suposa dezagun n ezaugarri edo aldagai ditugula. Binaka hartuz, kobariantzak eta koerlazio-koefizienteak atera daitezke; orduan, **L** eta **R** matrizeak hauexek izango lirateke:

$$\mathbf{L} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1n} \\ l_{21} & l_{22} & \dots & l_{2n} \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{bmatrix}$$

non $l_{11}, l_{22}, \dots l_{nn}$, n aldagaien bariantzak diren hurrenez hurren; $l_{1n} = l_{n1}$, n eta 1 aldagaien arteko kobariantza da.

$r_{1n} = r_{n1}$, n eta 1 aldagaien arteko koerlazio-koefizientea da eta diagonaleko 1 balioa aldagai bakoitzak berekiko duen koerlazioa da.

V. KOERLAZIOA ETA ERREGRESIOA

V.1. SARRERA

V.2. ERREGRESIOA R^2-n

V.2.1. Batezbestekoaren erregresioa

V.2.2. Karratu txikienen erregresioa

V.3. KARRATU TXIKIENEN ERREGRESIO LINEALA R^2-n

V.3.1. Karratu txikienen erregresio zuzena

V.3.2. Karratu txikienen erregresio linealaren propietateak

V.3.3. Hondar bariantza eta mugatze-koefizientea

V.3.4. Erregresio-koefizientearen eta koerlazio--koefizientearen zeinuaren azterketa. Doikuntzaren egokitasuna

V.4. ERREGRESIO LINEALA R^n-n

V.4.1. Sarrera

V.4.2. Erregresio hiperplanoa

V.4.3. Propietateak

V.4.4. Erregresio hiperplanoaren koefizienteak lortzeko metodoa

V.4.5. β erregresio partzial estandardizatuen koefizienteak

V.4.6. Hondar bariantza. Mugatze-koefizientea eta koerlazio-koefiziente anizkoitza

V.4.7. Edozein aldagai azalduaren erregresioaren orokorpena

V.4.8. Erregresio-koefiziente desberdinen zeinuaren azterketa. Doikuntzaren egokitasuna

V.5. KOERLAZIO PARTZIALA

V.5.1. Koerlazio partziala R^3-n

V.5.2. Koerlazio partziala R^n-n

V.5.2. Koerlazio partziala eta erregresio-koefizienteen arteko erlazioa

V.1. SARRERA

Bi edo aldagai gehiago aztertzen dituen analisi-estatistiko baten aurrean, horien elkar aldaketa oso garrantzitsua da, beraien arteko berdintasun edo desberdintasunak ez baitira kasualitatearen ondorioak izango, baizik eta gehienetan, beraien artean erlazio funtzionalaren ondorioak izango dira.

Elkar-aldaketa zehazteko ditugun teknikak, bi motatakoak dira:

1. Erregresio tekniken bidez, $Y = f(X_1, ..., X_n, a, b, ...)$ funtzio matematikoa lortuko dugu, zeinak, $X_1, X_2. ..., X_n$ aldagaien balioak ordezkatuz gero, Y-ren, aldagai dependentearen, balio teorikoak emango baitizkigu.

2. Koerlazio-tekniken bidez: aldagaien arteko kobariantza azalduko dizkigun koefiziente batzu lortuko ditugu.

V.2. ERREGRESIOA R^2-n

Maiztasun banaketa bikoitzean, aldagaiaren balio bakoitzari bestearen zenbait balio dagozkio. Hala ere, aldagai horietako bat independentetzat hartzen badugu eta erlazio matematikoa estimatzen badugu, berorren balio bakoitzari, aldagai dependentearen balio bakarra dagokio.

Funtzio matematiko hori optimizazio erizpide bati jarraituz lortuko dugu, dispertsio-diagrama edo puntu-hodeia kontutan harturik.

Erregresio-metodo asko daude:

1. Batezbestekoaren erregresioa.

2. Karratu txikienen erregresioa.
   Erregresio lineala.
   Erregresio parabolikoa.

Egin dugun erregresioaz, aldagai baten bidez bestea aurresateko baliatuko gara, hauxe baita, azken finean, aldagaien erlazioen azterketen helburua.

Edozein metodori jarraituz, bi erregresio desberdin kontsidera daitezke: Y-rena X-ekiko eta X-ena Y-rekiko.

V.2.1. Batezbestekoaren erregresioa

Y-ren X-ekiko erregresioa kontsideratuz, batezbestekoaren erregresioak aldagaiaren x_i i = 1, ..., k balio bakoitzari, (Y/X = x_i) banaketa baldintzatuaren $\overline{y}/x_i$ batezbesteko balioa elkartzen dio. Kasu honetan, x_i balio desberdinentzat, non i = 1, ..., k den, Y aldagaiaren batezbesteko balio baldintzatuak izango dira:

$$\overline{y} / x_1 = \sum_{j=1}^{l} y_j / x_1 f(y_i / x_1)$$

$$\overline{y} / x_2 = \sum_{j=1}^{l} y_j / x_2 f(y_i / x_2)$$

$$\vdots \quad = \quad \vdots$$

$$\overline{y} / x_k = \sum_{j=1}^{l} y_j / x_m f(y_i / x_k)$$

(x_i, $\overline{y}/x_i$), i = 1, ..., k puntuak asko eta hurbilak balira, kurba bat osatuko lukete. Kurba hori, batezbestekoaren erregresioarena edo erregresio enpirikoarena deritzo, hain zuzen.

V.2.2. Karratu txikienen erregresioa

(X eta Y) bi aldagai estatistikoen puntu-hodeiaren arabera, beraien artean Y = f(X, a, b) moduko erlazio funtzional bat dagoela pentsa dezakegu, horrela, X aldagai independentea edo azaltzailea izango litzateke eta Y, berriz, aldagai dependentea edo azaldua. Puntu-hodeiaren itxuraren arabera, f(X, a, b) funtzio bat edo beste hartuko dugu.

Karratu txikienen zentzuan, karratu txikienen erregresioaren arabera funtzio hoberena doituko dugu. Doikuntza hau, a eta b parametroak lortzean datza, non hauek puntu-hodeiari hobekien doitzen zaion kurbarenak diren. Hau da, hurrengo balioa minimo egiten duen kurba aukeratu nahi dugu.

$$\sum_{i=1}^{m} \left[y_j - f(x_i, a, b) \right]^2$$

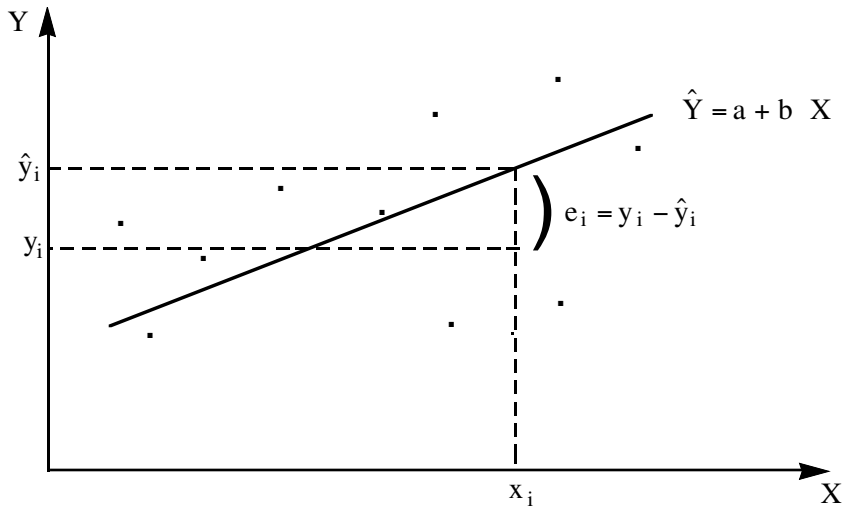
Kasu sinpleena funtzio lineala aukeratzen dugun kasua da, hau da, f(X, a,b) = a+ bX.

V.3. KARRATU TXIKIENEN ERREGRESIO LINEALA R^2-n

V.3.1. Karratu txikienen erregresio zuzena

(x_i, y_i), i = 1, ..., m puntu-hodeia emanik, eta $\hat{Y} = a + bX$ doikuntzaren funtzioa aukeraturik, x_i X-en balio bakoitzari, Y-ren bi balio dagozkio: puntu-hodeiko y_i balio erreala eta $X = x_i$ balioa aurreko zuzenean ordezkatuz lortzen dugun $\hat{y}_i$ balio teorikoa edo estimatua.

1 definizioa. y_i balio erreala eta $\hat{y}_i$ balio estimatuaren arteko diferentziari, errorea edo hondarra deritzo, eta e_i izendatuko dugu; hau da, $e_i = y_i - \hat{y}_i$, i = 1, ..., m.



Karratu txikienen eredua a eta b mugatzean datza, non erroreen karratuen batura minimo egiten duten.

Beraz, hurrengo problema askatu behar dugu:

$$\min_{a,b} \Phi(a,b) = \min_{a,b} \sum_{i=1}^{m} e_i^2 = \min_{a,b} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \min_{a,b} \sum_{i=1}^{m} (y_i - a - bx_i)^2 \qquad (5.1)$$

Φ funtzioak minimoa izateko beharrezko baldintza, a-rekiko eta b-rekiko deribatuak zero egitea da, hau da:

$$\frac{\partial \Phi}{\partial a} = 0 \Rightarrow ma + (\sum x_i)b = \sum y_i \qquad (5.2)$$

$$\frac{\partial \Phi}{\partial b} = 0 \Rightarrow (\sum x_i)a + (\sum x_i^2)b = \sum x_i y_i$$

non (5.2)-ko batukariak, eta baita aurrerantzean azalduko direnak,1-tik m-rainokoak diren. (5.2) sistemari **ekuazio normalen sistema** deritzogu.

(5.2) sistema askatuz, ondorengo balioak lortzen ditugu:

$$a = \bar{y} - b\bar{x} \qquad (5.3)$$

$$b = \frac{S_{xy}}{S_x^2}$$

non a, gai independentea den eta b, Y-ren X-ekiko erregresio zuzenaren X-en koefizientea, erregresio-koefizientea bezala ere ezagutzen dena.

(5.3)-n lortutako balioentzat Φ funtzioak minimoa izateko baldintza nahikoa betetzen du.

Ikus ezazu, (5.3)-n lortutako lehen ekuaziotik $\bar{y}$ askatuz, erregresio zuzena grabitate zentrutik pasatzen dela.

(5.3)-n lortutako a-ren eta b-ren balioak, Y = a+bX zuzenaren ekuazioan ordezkatuz, normalean erabiltzen den ondoko adierazpena lortzen dugu:

$$\hat{Y} - \bar{y} = \frac{S_{xy}}{S_x^2}(X - \bar{x})$$

Ondoren daukagun **X** matrizea

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}$$

eta $\mathbf{b} = (a,b)^T$; $y = (y_1,...,y_m)^T$ bektorea kontutan izanik, orduan zera daukagu:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} m & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \quad \text{eta} \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

eta horrela, (5.2) ekuazio normalen sistema matrizialki adieraz daiteke:

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y} \tag{5.4}$$

Baldin $det(\mathbf{X}^T\mathbf{X}) \neq 0$, aurreko sistemaren ebazpidea existitzen da eta izango da:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{5.5}$$

(5.5.) ebazpidea deribazio bektorialaren bitartez erraz lortzen da, ondoren ikusiko dugu bezala.

Izan bedi $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ erroreen bektorea, non $\hat{\mathbf{y}} = (\hat{y}_1,\ldots,\hat{y}_m)^T$ estimatutako balioen bektorea den. Horrela, (5.1) problema era honetan adierazten da:

$$\min \mathbf{e}^T\mathbf{e} \tag{5.6}$$

Estimatutako balioen bektorea $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ denez, (5.6) hurrengo adierazpenaren baliokidea izango da.

$$\min_{\mathbf{b}} (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) = \min_{\mathbf{b}} (\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}) \tag{5.7}$$

eta $\mathbf{b}$-rekiko bektorialki deribatuz, (ikus A.7 eranskina), zera lortzen da:

$$\frac{\partial \mathbf{e}^T\mathbf{e}}{\partial \mathbf{b}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{0} \tag{5.8}$$

nondik

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y} \tag{5.9}$$

lortzen den eta (5.4)-ren berdina den.

V.3.2. Karratu txikienen erregresio linealaren propietateak

Aurreko atalean egindako garapenetatik hurrengo propietateak ondoriozta daitezke:

i) Erroreen batezbestekoa zero da: $\bar{e} = 0$

Frogapena: (5.4)-tik edo (5.8)-tik zera daukagu:

$$\mathbf{X^T(y - Xb) = 0} \qquad (5.10)$$

non bi ekuazioko sistema den. Lehenengoak $\sum e_i = 0$ dela diosku, eta beraz, $\bar{e} = 0$.

ii) Estimatutako balioen batezbestekoa eta balio errealena berdinak dira: $\bar{\hat{y}} = \bar{y}$

Frogapena: **i)**-tik eta erroreen definiziotik berehala ondorioztatzen da.

iii) X aldagai azaltzailea eta e errore aldagaia elkarren artean koerlaziogabeak dira: $S_{xe} = 0$.

Frogapena: (5.10)-ko bigarren ekuaziotik, $\sum x_i e_i = 0$ lortzen dugu. Erroreek batezbestekoz zero dutenez, $S_{xe} = \frac{1}{m}\Sigma x_i e_i = 0$ da.

iv) $\hat{Y}$ **estimatutako aldagaia eta e errore aldagaia koerlaziogabeak dira**: $S_{\hat{y}e} = 0$

Frogapena: $S_{\hat{y}e} = S_{(a+bx)e} = bS_{xe} = 0$

v) Bariantzaren deskonposaketa batukorra: $S_y^2 = S_{\hat{y}}^2 + S_e^2$

Frogapena: $Y = \hat{Y} + e$ dela kontutan izanik, eta $(\hat{Y}, e)$ koerlaziogabeak direnez, baturaren bariantza bariantzen batura dela dakigu.

vi) $S_e^2 = (1 - r^2)S_y^2$

Frogapena (5.3)-ko b-ren adierazpenetatik eta $r = r_{xy} = \frac{S_{xy}}{S_x S_y}$ dela kontutan izanik, zera idatz dezakegu:

$$S_e^2 = S^2_{(y-\hat{y})} = S^2_{(y-a-bx)} = S^2_{(y-bx)} = S_y^2 + b^2S_x^2 - 2bS_{xy} = \tag{5.11}$$

$$= S_y^2 + \frac{S_{xy}^2}{S_x^2 S_x^2} S_x^2 - 2\frac{S_{xy}^2}{S_x^2} = S_x^2 - r^2 S_y^2 = (1 - r^2)S_y^2$$

vii) $S_{\hat{y}}^2 = r^2 S_y^2$

Frogapena: **v)** eta **vi)** propietateetatik berehala ondorioztatzen da.

V.3.3. Hondar bariantza eta mugatze-koefizientea

Orain oso erraza zaigu koerlazio-koefizientearen karratua 0 eta 1 tartean dagoela frogatzea. **vi)** propietatetik eta bariantzak ez negatiboak direla kontutan izanik, $(1 - r^2)$ ez negatiboa dela ondorioztatu dezakegu, eta beraz, lehen azaldutakoa. Gainera, koerlazio-koefizientea -1 eta 1 tartean dagoela frogatzen da.

2 definizioa. Koerlazio-koefizienteren karratuari *mugatze-koefizientea* deritzo.

vii) propietatean ikusi dugunez, mugatze-koefizientea azaldutako bariantza totalarekiko portzentaia bezala ulertarazi daiteke.

Ikus ezazu, X-en Y-rekiko erregresioa egin nahi badugu, garapena lehenengokoaren berdina dela, baina X eta Y aldagairen betebeharrak elkar aldatuz, hortaz, orain $\hat{X} = a + bY$.

Bereziki, erregresio honen b koefiziente berria $b = \frac{S_{xy}}{S_y^2}$ izango litzateke eta normalean erabiltzen den zuzenaren adierazpena:

$$(\hat{X} - \overline{x}) = \frac{S_{xy}}{S_y^2}(Y - \overline{y})$$

Bi erregresioak elkarrekin lantzen baditugu, b_{xy} X-en Y-rekiko erregresioaren $(\hat{X}(Y))$ koefizientea izango da eta b_{yx}, berriz, Y-ren X-ekikoa $(\hat{Y}(X))$.

Erregresio linealaren koefizientea erregresio zuzenaren malda da. Orduan, b_{yx}-ek X-en aldaketekin Y-ren gehikuntza-tasa neurtzen digu. Hau da, b_{yx}-ek X aldagaia unitate batetan handitzen denean Y aldagaiaren aldakuntza adierazten digu.

Berehala froga dezakegu, mugatze-koefizientea bi erregresio-koefizienteen biderkadura dela. Hau da:

$$r^2 = b_{xy}b_{yx} \tag{5.12}$$

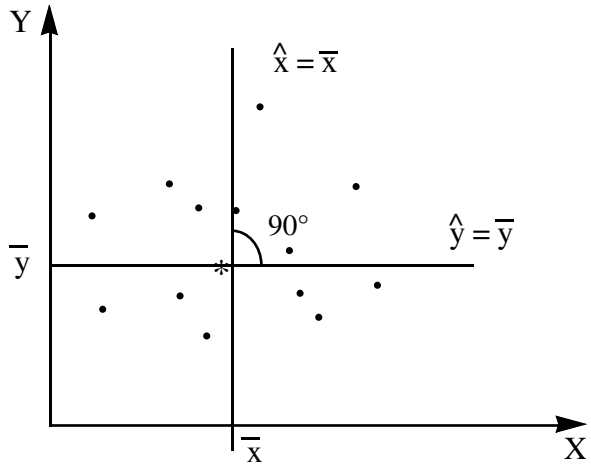
V.3.4. Erregresio-koefizientearen eta koerlazio-koefizientearen zeinuaren azterketa. Doikuntzaren egokitasuna

$$r_{xy} = r_{yx} = \frac{S_{xy}}{S_x S_y},\ b_{xy} = \frac{S_{xy}}{S_y^2},\ b_{yx} = \frac{S_{xy}}{S_x^2}$$

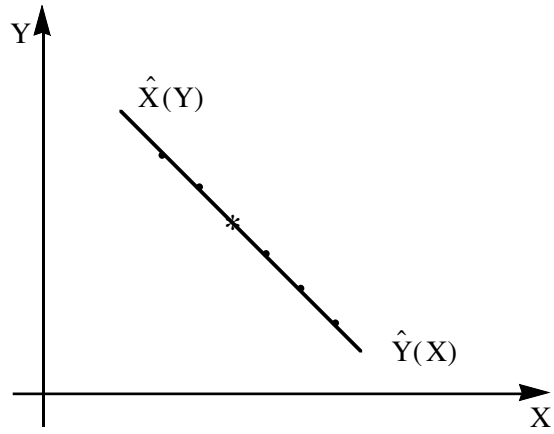
izanik eta bariantzak eta desbidazio tipikoak positiboak direnez, koefiziente hauek aldagaien arteko kobariantzaren zeinu berdina izango dute. Hau da, koerlazioa (eta beraz kobariantza) positiboa bada, bi zuzenen maldak positiboak lirateke. Analogoki izango da kobariantza negatiboa denean.

Koerlaziogabe direnen kasuan ($S_{xy} = 0$), lortutako zuzenak $\hat{Y} = \overline{y}$ eta $\hat{X} = \overline{x}$ izango dira, eta beraz, grafikoki ikus daitekeen bezala, ardatzen paraleloak izango dira eta beraien artean perpendikularrak. Koerlazio-koefizientea kontutan izanik, ondoren grafikoki aztertu ditzakegu:

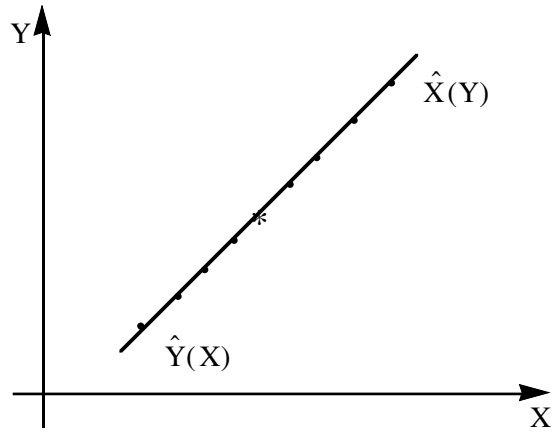
1. $r = 0$ ($S_{xy} = 0,\ S_{\hat{y}}^2 = 0$). Kasu honetan, zuzenen arteko angelua 90° da. X aldagaiak ez du linealki Y azaltzen eta beraz, azaldutako bariantza zero da.



2. $r = -1$ ($S_{xy} = -S_x S_y$ eta $S_e^2 = 0$). Bi zuzenek bat egiten dute, (osatutako angelua 0° da) eta malda negatiboa dute.



3. $r = 1$ ($S_{xy} = S_x S_y$ eta $S_e^2 = 0$). Bi zuzenek bat egiten dute, (osatutako angelua 0° da) eta malda positiboa dute.



Ikus dezakegunez, 2. eta 3. kasuetan koerlazioa osoa da eta hortaz, hodeiko puntu guztiak zuzenaren gainean daude, $y_i = \hat{y}_i$, $i = 1,\ldots,m$; kasu honetan doikuntza perfektua dela esango dugu.

Beraz, doikuntzaren egokitasuna r^2-ren balioarekin zehazta dezakegu. Berau 1-etik hurbil badago, doikuntza ona izango da, baina zerotik hurbil badago doikuntza txarra dela esango dugu.

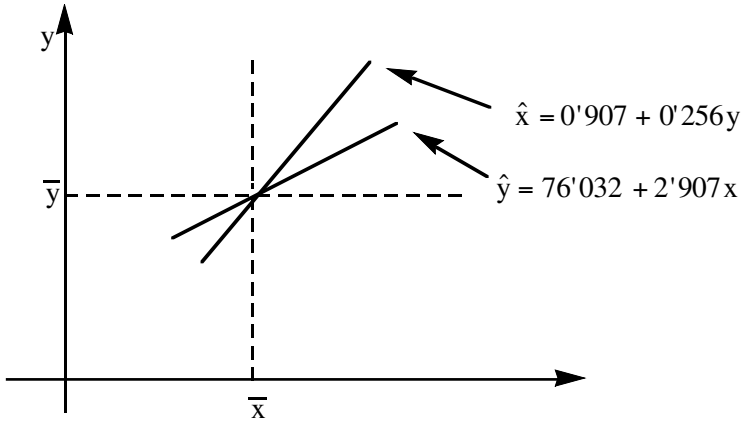
Adibidea:

X aldagaia lantegi batean egindako lan-orduak izanik eta Y aldagaia produzitu diren unitateak, kalkula dezagun koerlazio-koefizientea eta bi erregresio-zuzenak.

X	80	79	83	84	78	60	82	85	85	79	84
Y	300	302	315	330	300	250	300	340	315	330	310

Emaitzak hauek izan dira:

$r = 0'863$ $\quad \overline{x} = 79'9$ $\quad \overline{y} = 308'36$



Aurreko adibidean bi erregresio zuzenak egin baditugu ere, aldagaien arteko menpekotasuna alderantzizkoa da, hau da:

Bi galdera desberdin erantzuteko balio luketen zuzenak dira:

1) Zenbat lan-ordu beharko lirateke produzitutako unitateak kopuru jakina izateko?

2) Zein izango litzateke produzitutako unitateen kopurua egindako lan-orduak kopuru jakina bada?

V.4. ERREGRESIO LINEALA R^n-n

V.4.1. Sarrera

Atal honetan karratu txikienen erregresio linealaren analisia n aldagaietara zabalduko dugu $(X_1,...,X_n)$.

Kasu honetan, m indibiduoentzako n dimentsiotako banaketa baten oharpenak ditugu, non x_{ij}–k i indibiduoaren ohartutako balioa j aldagaiarako adierazten duen.

Aldagai baten (azaldua) beste n-1 aldagaiekiko (azaltzaileak) karratu txikienen erregresio lineala egin nahi dugu. Orokortasunik galdu gabe, X_1 aldagaia azaldutzat hartuko dugu eta

$\hat{X}_1(X_2,\ldots,X_n)$ erregresio hiperplanoa doituko dugu:

$$\hat{X}_1 = a_1 + b_{12}X_2 + \ldots + b_{1n}X_n \tag{5.13}$$

Ondorengo kalkuluak errazteko, R^2-n erabilitako notazioa zabalduko dugu. Horrela, **X** matrizea osatuko dugu:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1n} \\ 1 & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

X, beraz, aldagai independente bakoitzaren m obserbazioz osaturik dago, gehi 1-ez osatutako zutabe bat.

$$\mathbf{b} = (a_1, b_{12},\ldots, b_{1n})^T, \quad \mathbf{e} = \mathbf{x}_1 - \hat{\mathbf{x}}_1, \tag{5.14}$$

non $\mathbf{x}_1 = (x_{11},\ldots,x_{m1})^T$ X_1 aldagaiaren balio errealen bektorea den eta $\hat{\mathbf{x}}_1 = (\hat{x}_{11},\ldots,\hat{x}_{m1})^T$, X_1 aldagaiaren estimatutako bektorea; beraz, $\hat{\mathbf{x}}_1 = \mathbf{Xb}$.

Hurrengo bektorea, $\bar{\mathbf{x}}$, batezbestekoen bektore bezala definituko dugu:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{pmatrix} \tag{5.15}$$

eta $\mathbf{L}$, $(X_1,...,X_n)$ aldagaien arteko bariantza eta kobariantza matrizea izango da:

$$\mathbf{L} = \begin{pmatrix} S_1^2 & S_{12} & \cdots & S_{1n} \\ S_{12} & S_2^2 & \cdots & S_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ S_{1n} & S_{2n} & \cdots & S_n^2 \end{pmatrix} \tag{5.16}$$

non $S_{kj} = S_{x_k x_j} = \mathrm{Cov}(x_k, x_j)$ $k \neq j$, eta $S_j^2 = S_{x_j}^2 = \mathrm{Var}(x_j)$, non $j = 1,\ldots,n$.
Azkenik, **R** koerlazio-matrizea izango da:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{12} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{1n} & r_{2n} & \cdots & 1 \end{pmatrix} \tag{5.17}$$

non $r_{kj} = r_{x_k x_j} = \mathrm{Corr}(x_k, x_j) = \dfrac{S_{kj}}{S_k S_j}$ $k \neq j$.

L eta **R** matrize simetriko eta erdidefinitu positiboak dira, aurrerantzean erabiliko ditugun propietateekin (ikus V.A eranskina).

V.4.2. Erregresio hiperplanoa

(5.13) motako erregresio hiperplanoa doitzeko, karratu txikienen ereduaren arabera, R^2-n egindako minimizazio problema baliokidea planteatzen dugu matrizialki:

$$\min_{\mathbf{b}} \mathbf{e}^T\mathbf{e} = \min_{\mathbf{b}} (\mathbf{x}_1 - \mathbf{X}\mathbf{b})^T(\mathbf{x}_1 - \mathbf{X}\mathbf{b}) \tag{5.18}$$

non $\mathbf{e}$, $\mathbf{x}_1$, $\mathbf{b}$ eta $\mathbf{X}$ aurretik definitutakoak diren.

(5.18)-n **b**-rekiko bektorialki deribatuz eta zerori berdinduz, hurrengo ekuazio normalen sistema lortzen dugu:

$$\mathbf{X^T X b = X^T x_1} \tag{5.19}$$

eta $\det(\mathbf{X^T X}) \neq 0$ bada, (5.18) askatzen duen $\mathbf{b}$ parametroen bektorea izango da:

$$\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T x_1} \tag{5.20}$$

V.4.3. Propietateak

Orokorrean, n aldagaien kasuan, hurrengo erregresio propietateak ditugu:

i) Erroreak aldagai zentratuak dira: $\bar{e} = 0$.

Frogapena: Bi aldagaien kasuan bezala, n aldagaiekin (5.19)-tik zera daukagu:

$$\mathbf{X^T(x_1 - Xb) = X^T e = 0} \tag{5.21}$$

n ekuazio dituen sistema honen lehen ekuazioak $\sum e_i = 0$ dela diosku eta hemendik propietatea lortzen dugu.

ii) Balio estimatuen batezbestekoa eta balio errealenak berdinak dira: $\bar{\hat{x}}_1 = \bar{x}_1$

Frogapena: **i)** -tik eta erroreen definiziotik ondoriozta daiteke.

iii) Erroreak eta edozein aldagai azaltzaile elkarren artean koerlaziogabeak dira: $S_{x_j e} = 0$, $\forall j=2,...,n$.

Frogapena: (5.21)-ko beste n–1 ekuazioetatik ondorioztatzen da.

iv) Erroreak estimatutako balioekin koerlaziogabeak dira: $S_{\hat{x}_1 e} = 0$

Frogapena: $\hat{X}_1$, X_2, ..., X_n aldagaien konbinazio lineala dela jakinik, garbi ikus dezakegu.

v) Bariantzaren deskonposaketa batukorra: $S^2_{x_1} = S^2_{\hat{x}_1} + S^2_e$

Frogapena: $X_1 = \hat{X}_1 + e$ dela kontutan izanik eta $(\hat{X}_1, e)$ aldagaiak koerlaziogabeak direnez, propietatea lortzen da.

V.4.4. Erregresio hiperplanoaren koefizienteak lortzeko metodoa

ii) propietatetik zera daukagu:

$$a_1 = \bar{x}_1 - b_{12}\bar{x}_2 - \ldots - b_{1n}\bar{x}_n \tag{5.22}$$

eta (5.14) hiperplanoa horrela adieraz dezakegu:

$$\hat{X}_1 - \bar{x}_1 = b_{12}(X_2 - \bar{x}_2) + \ldots + b_{1n}(X_n - \bar{x}_n) \tag{5.23}$$

R^2-n bezalaxe, lehendabizi hiperplanoa batezbestekoen bektoretik pasatzen dela ikusten da, hau da grabitate zentrutik. Bigarrenik, aldagaiak zentratuta daudela kontsideratzen badugu, erregresio-koefizienteak ez dira aldatzen. Orduan, orokortasunik galdu gabe, azken kasu honetarako koefizienteak kalkulatuko ditugu.

Aldagai zentratuentzat minimizazio prozesuaren ekuazio normalen sistema garatzen badugu, eta kasu honetan, honako matrize hauek

$$\mathbf{b} = (b_{12},\ldots,b_{1n})^T, \quad \mathbf{e} = \mathbf{x}_1 - \hat{\mathbf{x}}_1, \tag{5.24}$$

$$\mathbf{X} = \begin{pmatrix} x_{12} & \cdots & x_{1n} \\ x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots \\ x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

ditugularik, zera lortzen da:

$$\begin{aligned}
b_{12}\sum x_{i2}^2 &+ b_{13}\sum x_{i2}x_{i3} + \cdots + b_{1n}\sum x_{i2}x_{in} = \sum x_{i1}x_{i2} \\
b_{12}\sum x_{i2}x_{i3} &+ b_{13}\sum x_{i3}^2 + \cdots + b_{1n}\sum x_{i3}x_{in} = \sum x_{i1}x_{i3} \\
\vdots \quad & \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
b_{12}\sum x_{i2}x_{in} &+ b_{13}\sum x_{i3}x_{in} + \cdots + b_{1n}\sum x_{in}^2 = \sum x_{i1}x_{in}
\end{aligned} \tag{5.25}$$

x_{ij} bakoitzak i indibiduoaren balioa j aldagai zentraturako azaltzen duenez, (5.25) sistemako ekuazio bakoitza m-gatik zatitzen badugu, bariantza eta kobariantza terminotan adieraz dezakegu:

$$\begin{aligned}
b_{12}S_2^2 &+ b_{13}S_{23} + \cdots + b_{1n}S_{2n} = S_{12} \\
b_{12}S_{23} &+ b_{13}S_3^2 + \cdots + b_{1n}S_{3n} = S_{13} \\
\vdots \quad &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
b_{12}S_{2n} &+ b_{13}S_{3n} + \cdots + b_{1n}S_n^2 = S_{1n}
\end{aligned} \tag{5.26}$$

sistema matrizialki horrela gelditzen da:

$$\begin{pmatrix} S_2^2 & S_{23} & \cdots & S_{2n} \\ S_{23} & S_3^2 & \cdots & S_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ S_{2n} & S_{3n} & \cdots & S_n^2 \end{pmatrix} \begin{pmatrix} b_{12} \\ b_{13} \\ \vdots \\ b_{1n} \end{pmatrix} = \begin{pmatrix} S_{12} \\ S_{13} \\ \vdots \\ S_{1n} \end{pmatrix} \tag{5.27}$$

Cramer-en bidez askatzen badugu, b_{12} koefizientea izango da:

$$b_{12} = \frac{\begin{vmatrix} S_{12} & S_{23} & \cdots & S_{2n} \\ S_{13} & S_3^2 & \cdots & S_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ S_{1n} & S_{3n} & \cdots & S_n^2 \end{vmatrix}}{\begin{vmatrix} S_2^2 & S_{23} & \cdots & S_{2n} \\ S_{23} & S_3^2 & \cdots & S_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ S_{2n} & S_{3n} & \cdots & S_n^2 \end{vmatrix}} = -\frac{\mathbf{L}_{12}}{\mathbf{L}_{11}} \tag{5.28}$$

non $\mathbf{L}_{ij}$-k, **L** bariantza eta kobariantza matrizearen (i,j) adjuntua adierazten duen.

Analogoki, edozein b_{1j} koefizientearen adierazpena lor dezakegu:

$$b_{1j} = -\frac{\mathbf{L}_{1j}}{\mathbf{L}_{11}} \qquad j = 2,\ldots,n \tag{5.29}$$

L eta **R** matrizeen adjuntuen arteko erlazioa kontutan izanik, (ikusi V.A.4 eranskina), aurreko adierazpena **R** koerlazio-matrizearen adjuntuen bitartez adieraz dezakegu.

$$b_{1j} = -\frac{S_1 \mathbf{R}_{1j}}{S_j \mathbf{R}_{11}} \qquad j = 2,\ldots,n \tag{5.30}$$

Azkenik, koefizienteen balio guztiak ordezkatuz lortzen dugun erregresio hiperplanoaren ekuazio orokorra ondokoa da:

$$\hat{X}_1 - \bar{x}_1 = -\frac{\mathbf{L}_{12}}{\mathbf{L}_{11}}(X_2 - \bar{x}_2) - \ \ldots\ - \frac{\mathbf{L}_{1n}}{\mathbf{L}_{11}}(X_n - \bar{x}_n). \tag{5.31}$$

edota

$$\hat{X}_1 - \bar{x}_1 = -\frac{S_1 \mathbf{R}_{12}}{S_2 \mathbf{R}_{11}}(X_2 - \bar{x}_2) - \ \ldots\ - \frac{S_1 \mathbf{R}_{1n}}{S_n \mathbf{R}_{11}}(X_n - \bar{x}_n). \tag{5.32}$$

V.4.5. β erregresio partzial estandardizatuen koefizienteak

Aurretik aldagaiak tipifikatu izan bagenitu, erregresio hiperplanoa horrela adieraziko litzateke:

$$\hat{T}_1 = \beta_{12} T_2 + \ \ldots\ + \beta_{1n} T_n \tag{5.33}$$

non T_j, X_j hasierako aldagaiari dagokion aldagai tipifikatua den, eta β_{1j} notazioa, horrelako aldagaien erregresio-koefizienteentzako erabiliko dugun. Konturatu behar da, ez dagoela gai independenterik, aldagai tipifikatuak zentratuak baitaude. Gainera, aldagai hauek bariantzaz 1 dute, eta beraz, (5.30) erabiliz, erregresio--koefizienteak kasu honetan izango dira:

$$\beta_{1j} = -\frac{\mathbf{R}_{1j}}{\mathbf{R}_{11}} = \frac{b_{1j}}{S_1 / S_j} \qquad j = 2,\ldots,n. \tag{5.34}$$

3 definizioa. (5.34)-n lortutako koefizienteak *erregresio partzial standardizatuen koefizienteak* dira.

Oharra: R^2-ren erregresioan, erregresio partzial standardizatuen koefizientea koerlazio-koefizientea $\beta_{12} = \beta_{21} = r_{12}$ da.

Koefiziente hauen garrantzia, adibide praktikoetan izango ditugun erantzunen interpretaziotan ikusiko da.

V.4.6. Hondar bariantza. Mugatze-koefizientea eta koerlazio-koefiziente anizkoitza

Erroreen batezbestekoa zero denez, bere bariantza matrizialki horrela adieraz dezakegu:

$$S_e^2 = \frac{1}{m}\mathbf{e}^T\mathbf{e} = \frac{1}{m}\left[\mathbf{x}_1 - \mathbf{Xb}\right]^T\left[\mathbf{x}_1 - \mathbf{Xb}\right]. \tag{5.35}$$

Aurreko biderkaketa garatuz eta **b**, (5.19)-ko balioaz ordezkatuz, (5.35) hurrengoaren baliokidea da:

$$S_e^2 = \frac{1}{m}\left[\mathbf{x}_1^T\mathbf{x}_1 - \mathbf{x}_1^T\mathbf{Xb}\right] \tag{5.36}$$

Ondorengo eragiketak errazteko, aldagaiak zentratuak daudela suposa dezakegu. Kasu orokorrari zabaltzea erraza izango litzateke, zeren badakigu aldagaien traslazioek hondar bariantzan ez dutela eragiten.

Horrela, datozen balio hauek (5.36)-ean ordezkatzen baditugu:

$$\mathbf{x}_1^T\mathbf{x}_1 = mS_1^2$$

$$\mathbf{x}_1^T\mathbf{X} = (mS_{12} \ \ldots \ mS_{1n}) \quad \text{eta} \tag{5.37}$$

$$\mathbf{b} = \left(-\frac{\mathbf{L}_{12}}{\mathbf{L}_{11}} - \frac{\mathbf{L}_{13}}{\mathbf{L}_{11}} \ \ldots \ -\frac{\mathbf{L}_{1n}}{\mathbf{L}_{11}}\right)^T$$

$\mathbf{L}_{11} \neq 0$ bada, hondar bariantzaren adierazpena lortzen dugu:

$$S_e^2 = S_1^2\frac{\mathbf{L}_{11}}{\mathbf{L}_{11}} + S_{12}\frac{\mathbf{L}_{12}}{\mathbf{L}_{11}} + \ \ldots \ + S_{1n}\frac{\mathbf{L}_{1n}}{\mathbf{L}_{11}} = \frac{|\mathbf{L}|}{\mathbf{L}_{11}} = \frac{|\mathbf{R}|}{\mathbf{R}_{11}}S_1^2 \tag{5.38}$$

non $|\mathbf{L}|$ eta $|\mathbf{R}|$, $\mathbf{L}$ eta $\mathbf{R}$ matrizeen determinanteak diren.

1 oharra. $\mathbf{L}$ eta $\mathbf{R}$ matrize erdidefinitu positiboak direnez, $|\mathbf{L}|$, $\mathbf{L}_{11}$ eta $|\mathbf{R}|$, $\mathbf{R}_{11}$ ezin dira negatiboak izan, eta hau hondar bariantzaren definizioarekin ados dago.

Mugatze-koefiziente anizkoitza

4 definizioa. X_1-en X_2, ..., X_n -rekiko erregresioaren $r^2_{1\bullet 2,\ldots,n}$ *mugatze-koefiziente anizkoitza*, azaldutako bariantzaren aldagai azalduaren bariantza totalarekiko proportzioa da.

Analitikoki,

$$r^2_{1\bullet 2\cdots n} = \frac{S^2_{\hat{x}_1}}{S^2_{x_1}} = 1 - \frac{S^2_e}{S^2_{x_1}}, \tag{5.39}$$

Definizio honetatik, R^2-n lortutako **vi** propietatearen orokorpena R^n-rako lortzen da.

vi) $S^2_e = \left(1 - r^2_{1\bullet 2\cdots n}\right)\ S^2_1,$

eta beraz, $0 \le r^2_{1\bullet 2\cdots n} \le 1$ ere betetzen da.

(5.38)-tik, (5.39) horrela adieraz daiteke:

$$r^2_{1\bullet 2\cdots n} = 1 - \frac{|\mathbf{L}|}{\mathbf{L}_{11}S^2_1} = 1 - \frac{|\mathbf{R}|}{\mathbf{R}_{11}} \tag{5.40}$$

R-ren determinantea bere lehenengo lerroko adjuntuetaz garatzen badugu, $|\mathbf{R}| = \mathbf{R}_{11} + r_{12}\,\mathbf{R}_{12} + \ldots + r_{1n}\,\mathbf{R}_{1n}$, mugatze-koefizientea standardizatuak erregresio--koefizienteen eta koerlazio koefizienteen arteko biderkaduren batura izango da.

$$r^2_{1\bullet 2\cdots n} = \beta_{12}r_{12} + \beta_{13}r_{13} + \ldots + \beta_{1n}r_{1n} \tag{5.41}$$

5 definizioa. Mugatze-koefiziente anizkoitzaren erro karratu positiboari koerlazio-koefiziente anizkoitza deritzogu eta $r_{1\bullet 2\cdots n}$ bezala adierazten dugu.

Beraz,

$$r^2_{1\bullet 2\cdots n} = \sqrt{1 - \frac{S^2_e}{S^2_1}} = \sqrt{1 - \frac{|\mathbf{L}|}{\mathbf{L}_{11}S^2_1}} \tag{5.42}$$

Koerlazio-koefiziente anizkoitza, bi aldagaien arteko koerlazio sinple edo totalaren arabera azaldu daiteke eta X_1 aldagai azalduaren eta beste aldagai azaltzaile guztien (X_2, ... ,X_n) arteko erlazioa azaltzen digu.

Bi aldagai genituenean, koerlazioak zuen zeinuaz hitzegiteak zentzua zuen, aldagaien kobariantza zuzena edo alderantzizkoa azaltzen baitzuen. Dimentsio haundiagoak ditugunean berriz, azaltzaile batek eragin positiboa izan dezake eta beste batzuk ordea, negatiboa, eta beraz, koerlazio-koefizientearen zeinuaren justifikaziorik ez dago.

Baina $\hat{X}_1$ aldagai azaltzaileen konbinazio lineala denez, $r_{1\bullet 2\cdots n}$ X_1 eta $\hat{X}_1$-ren arteko koerlazio totalatzat har dezakegu, eta horrela, zeinu positiboa baldin badu, aldagai baten eta bere estimatuaren arteko koerlazioa positiboa dela adieraziko du.

V.4.7. Edozein aldagai azalduaren erregresioaren orokorpena

Aldagai azaldua X_j, j=1, ...,n , edozein aldagai denean, formulak aurrekoen berdinak dira, baina aldagai independentearen eta dependentearen betebeharrak era egokian aldatu behar dira. Horrela, orokorrean,

$$\hat{X}_j - \overline{x}_j = \sum_{k \neq j} b_{jk}\left(X_k - \overline{x}_k\right), \qquad \text{non} \qquad b_{jk} = -\frac{\mathbf{L}_{jk}}{\mathbf{L}_{jj}} \tag{5.43}$$

$$\hat{T}_j = \sum_{k \neq j} \beta_{jk} T_k, \qquad \text{non} \qquad \beta_{jk} = -\frac{\mathbf{R}_{jk}}{\mathbf{R}_{jj}} \quad \text{eta,} \tag{5.44}$$

$$r^2_{1\bullet 2\cdots(j-1)(j+1)\cdots n} = 1 - \frac{|\mathbf{L}|}{\mathbf{L}_{jj} S_j^2} \tag{5.45}$$

Garbi dago, erregresioaren propietateak berridazten direla, aldagai azaltzaile desberdinei notazioa egokituz.

V.4.8. Erregresio-koefiziente desberdinen zeinuaren azterketa. Doikuntzaren egokitasuna

X_1-ek azaldutako aldagaia izaten jarraituko du. Lehendabizi, (5.30) eta (5.34) parekatuz, b_{1j} eta β_{1j}-ren zeinuak, $\mathbf{R}_{1j}$-ren zeinuaren araberan daudela ikus dezakegu, non biek zeinu berbera izango duten.

(5-23)-ko hiperplanoaren ekuazioan dakusagunez, x_j unitate batean gehitzen denean, beste aldagai azaltzaileak konstante mantenduz, X_1 aldagaiaren hazkuntza edo murriztapena b_{1j} baliokoa izango da. Partikularki, b_{1j} koefizienteen zeinuak bi aldagai hauen kobariantza, zuzena edo alderantzizkoa, erakusten du, besteak konstante mantenduz.

Aldagai tipifikatuekin lan eginez gero, β_{1j} koefizientea jatorrizko aldagaien neurri unitateekiko independentea izango da, baina bere zeinuaren esanahia b_{1j}-renaren berdina, honek aurretik azaldutako zeinuen berdintasuna justifikatzen duelarik. Bestalde, β koefiziente standardizatuen neurri unitateekiko independenteak izatean, beraien artean gonbaragarriak dira eta azaltzaileek azalduaren gain duten eragin erlatiboaren adierazle bezala erabiltzen dira, koefiziente standardizatuaren modulu haundiagoa duena eraginkorragoa delarik.

Bi aldagaiekin bezala, hiperplanoaren datuekiko doikuntzaren egokitasunaren analisia egin nahi dugu, eta mugatze-koefizientea adierazle egokia da. Lehen bezala, $r^2_{1\bullet 2,\ldots,n}$-ren balioa 1 -etik hurbil badago, doikuntza oso ona izango da $(S_e^2 \approx 0)$ eta 0 -tik hurbil badago, berriz, doikuntza txarra izango dugu $(S_e^2 \approx S_1^2)$, 0 eta 1 har ditzakeen balioen muturrak izanik.

V.5. KOERLAZIO PARTZIALA

Har dezagun Jamaikako ron-aren prezioa eta apaizen alokairua bi aldagai estatistikoen elkar eboluzioa. Edozein arrazoi izanik, bi aldagaien arteko koerlazio-koefizientea kalkulatu nahi dugu, eta seguruenik balio positiboa eta altua izango da. Honek bien artean dependentzia dagoela esan nahi al du?

Inongo ekonomilariari ez litzaioke bururatuko Jamaikako ron-aren prezioa jeisteko, apaizen alokairua jeitsi behar dugunik!

Honek, aldagaiekin zerikusi zuzenik ez duten kausek eragiten dituzten koerlazioak eta koerlazio kausalak, hau da, aldagaien arteko dependentzia garbia azaltzen duten erlazioak-bereizi behar direla esan nahi du. Zentzu honetan, emandako adibidearen koerlazioak lehen motako koerlazioa azaltzen digu, hau da, beste hirugarren aldagai baten (edo gehiagoren) eraginagaitik eman dela, inflazioa adibidez. Aldagai honek, bere aldetik, beste bi aldagaiekin koerlazio kausal bat izan dezake. Hau da, inflazioa handitzen bada, prezioak eta alokairuak batera handitzea izan daiteke.

Beraz, atal honetan beste aldagaien eragina kontrolatzen duen koerlazio kontzeptu baten definizio bat eman nahi dugu.

Har dezagun beste adibide berri bat erreferentziatzat.

Amerikako zenbait herriren kriminaltasun-tasaren kausak aztertu nahi ditu ikertzaile batek. Horretarako, hurrengo lau aldagaien datuak hartu ditu:

X_1 : koloreko populazioaren portzentaia.
X_2 : 3000 dolar baino gutxiagoko sarrera duten populazioaren portzentaia.
X_3 : herriaren tamainua.
X_4 : kriminaltasun-tasa.

Datu horiekin ondorengo koerlazio-matrizea lortu da:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.51 & 0.41 & 0.36 \\ & 1 & 0.29 & 0.60 \\ & & 1 & 0.49 \\ & & & 1 \end{pmatrix} \qquad (5.46)$$

Aurreko matrizean (X_4) kriminaltasun-tasa eta (X_1) koloreko populazioaren portzentaiaren arteko koerlazioa, nahiz eta oso haundia ez izan, positiboa dela ikusten da. (X_2) pobreziak eta (X_3) herriaren masifikazioak X_4-rekin ere koerlazio positiboa eta haundiagoa dute, eta era berean, aldagai horiek X_1-ekin ere koerlazio positiboa dute.

Ikertzaileak kriminaltasun eta koloreko populazioaren arteko koerlazio totala benetakoa den edota X_2 eta X_3 bezalako beste aldagaien eraginagaitik den galdetu behar du. Horretarako, azken aldagai hauen eragina baztertu behar du eta ondoren, koerlazioen aldaketak aztertu. Aldagai batek beste bien arteko erlazioan duen eragina baztertzeko, baztertu nahi duguna aldaketarik ez lukeen ingurune batean kokatuz lortuko genuke, hau da, aldagai hori konstante egongo balitz. Ingurune hau sortzea ikerketa gehienetan posible ez denez koerlazio partziala zer den aztertuko dugu.

V.5.1. Koerlazio partziala R^3-n

6 definizioa. X_1 eta X_2-ren arteko *koerlazio partziala* X_3-ren eragina kenduta, $r_{12 \bullet 3}$ deritzogu. Koerlazioa, X_1-en X_3 aldagaiarekiko eta X_2-ren X_3 aldagaiarekiko erregresioetan sortzen diren erroreen arteko koerlazio totala da.

Hau da,

$$r_{12\bullet 3} = \mathrm{Corr}\left(e_{1\bullet 3},\ e_{2\bullet 3}\right) = \frac{\mathrm{Cov}\left(e_{1\bullet 3},\ e_{2\bullet 3}\right)}{\sqrt{\mathrm{Var}\left(e_{1\bullet 3}\right)}\sqrt{\mathrm{Var}\left(e_{2\bullet 3}\right)}} \tag{5.47}$$

non

$$\begin{aligned} e_{1\bullet 3} &= X_1 - \hat{X}_1(X_3) = (X_1 - \bar{x}_1) - \frac{s_{13}}{s_3^2}(X_3 - \bar{x}_3) \\ e_{2\bullet 3} &= X_2 - \hat{X}_2(X_3) = (X_2 - \bar{x}_2) - \frac{s_{23}}{s_3^2}(X_3 - \bar{x}_3) \end{aligned} \tag{5.48}$$

Beraz, horrela definituriko koerlazio-koefiziente partzialak X_1 eta X_2-ren arteko koerlazio linealaren maila neurtzen du, baina bietan X_3-ren eragin lineala kendu ondoren.

(5.47) definizioa, r_{12}, r_{13}, r_{23} koerlazio totalen arabera azal dezakegu. Horretarako, aldagaien transformazio lineal zuzena egiten badugu, koerlazio--koefizientea ez dela aldatzen gogoratu behar dugu. Bereziki, aldagaiak tipifikaturik daudela pentsa dezakegu. Kasu honetan, (5.48)-ko adierazpenak hurrengoak izango dira:

$$\begin{aligned} e_{1\bullet 3} &= T_1 - r_{13}T_3 \\ e_{2\bullet 3} &= T_2 - r_{23}T_3 \end{aligned} \tag{5.49}$$

Horrela,

$$\begin{aligned} \mathrm{Cov}(e_{1\bullet 3}, e_{2\bullet 3}) &= \mathrm{Cov}(t_1 - r_{13}t_3,\ t_2 - r_{23}t_3) = \\ &= r_{12} - r_{13}r_{23} - r_{23}r_{13} + r_{13}r_{23} = r_{12} - r_{13}r_{23} \end{aligned} \tag{5.50}$$

$$\mathrm{Var}(e_{1\bullet 3}) = \mathrm{Var}(t_1 - r_{13}t_3) = 1 - r_{13}^2 \tag{5.51}$$

$$\mathrm{Var}(e_{2\bullet 3}) = \mathrm{Var}(t_2 - r_{23}t_3) = 1 - r_{23}^2 \tag{5.52}$$

Azkenik, (5.50), (5.51) eta (5.52) ordezkatuz (5.47)-an, hauxe lortzen dugu:

$$r_{12\bullet 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \tag{5.53}$$

(5.53) erabiliz, eta gure adibideko aldagaiei aplikatuz, X_1 eta X_4 aldagaien arteko koerlazioa azter dezakegu, beste aldagaiek bakoitzean duten eragina baztertuz.

$$r_{14\bullet 2} = \frac{r_{14} - r_{12}r_{24}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{24}^2}} = \frac{0.36 - 0.60 \times 0.51}{\sqrt{1 - (0.60)^2}\sqrt{1 - (0.51)^2}} = 0.078$$

$$r_{14\bullet 3} = \frac{r_{14} - r_{13}r_{34}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{34}^2}} = \frac{0.36 - 0.41 \times 0.49}{\sqrt{1 - (0.41)^2}\sqrt{1 - (0.49)^2}} = 0.2$$

Erantzun hauekin, ikertzaileak $r_{14} = 0.36$ eta $r_{14\bullet 2} = 0.078$-ren arteko diferentzia aztertzen du, eta koloreko populazioaren eta kriminaltasun-tasaren arteko koerlazio totalaren zati handiena sarrera txikiko populazioaren portzentaiak bi aldagaietan duen eraginagaitik dela ondoriozta dezake.

Analisi honekin jarraituz, ikertzaileak bi aldagaien eragina batera atean kentzea pentsatzen du, pobrezia eta masifikazioa.

V.5.2. Koerlazio partziala R^n-n

7 definizioa. X_1 eta X_2 aldagaien arteko *koerlazio partzialari,* X_3, ..., X_n aldagaien eragina baztertuz, $r_{12\bullet 3\cdots n}$ deritzogu, eta $r_{12\bullet 3\cdots n}$ koerlazioak X_1-en eta X_2-ren X_3, ...,X_n-rekiko erregresioetan sortzen diren hondarren arteko koerlazioa da.

Hau da:

$$r_{12\bullet 3\cdots n} = \mathrm{Corr}(e_{1\bullet 3\cdots n}, e_{2\bullet 3\cdots n}) \tag{5.54}$$

X_1 eta X_2 aldagaien arteko koerlazio ohartuan, X_3 eta X_4 aldagaien eragina baztertzeko erabiltzen edo jarraitzen den prozesua, hauxe da: lehen horietatik batena baztertzea, eta ondoren lehenengo erregresioan lortutako erroreen bestearen eragina baztertzea baliokidea da. Hau errepikapen eredua da.

Logikoa denez, prozesu honen emaitza honakoa da:

$$r_{12\bullet 34} = \frac{r_{12\bullet 3} - r_{14\bullet 3}r_{24\bullet 3}}{\sqrt{1 - r_{14\bullet 3}^2}\sqrt{1 - r_{24\bullet 3}^2}} \tag{5.55}$$

Noski, emaitza berdina izango litzateke, lehendabizi X_4-ren eragina

baztertuko bagenu; hau da:

$$r_{12\bullet 34} = \frac{r_{12\bullet 4} - r_{13\bullet 4} r_{23\bullet 4}}{\sqrt{1 - r^2_{13\bullet 4}}\sqrt{1 - r^2_{23\bullet 4}}} \tag{5.56}$$

Errepikapen eredu hau, erraza izan arren, aldagaien kopurua handitzen denean lankorra bihurtu daiteke.

Horregatik, **L** edo **R**-ren adjuntuen bitartez, baliokidea den matrizeen adierazpen bat erabiltzen da.

Orokorrean, $r_{12\bullet 3\cdots n}$, X_1 eta X_2 aldagaien arteko koerlazioa bada, beste n-2 aldagaien eragina kenduta, zera daukagu:

$$r_{12\bullet 3\cdots n} = -\frac{\mathbf{L}_{12}}{\mathbf{L}_{11}^{1/2}\mathbf{L}_{22}^{1/2}} = -\frac{\mathbf{R}_{12}}{\mathbf{R}_{11}^{1/2}\mathbf{R}_{22}^{1/2}} \tag{5.57}$$

Adibidearekin jarraituz,

$$r_{14\bullet 23} = -\frac{\mathbf{R}_{14}}{\mathbf{R}_{11}^{1/2}\mathbf{R}_{44}^{1/2}} = -0.061$$

non **R** matrizea, X_1, X_2, X_3 eta X_4 aldagaientzat (5.46)-ean azaldutako koerlazio--matrizea den.

Horrela, koloreko populazioaren portzentaia eta kriminaltasun-tasaren arteko koerlazio sinple positiboa, aldagaien arteko zuzeneko elkar dependentziagaitik **ez** dela ikusten dugu, baizik eta masifikazio edota pobreziak bi aldagaietan duten eragin zuzenagaitik izan daiteke.

Erantzun hauek direla eta, zera pentsa dezakegu: beharbada pobrezia eta masifikazio maila berdina duten populazioekin beste ikerketa paralelo bat egingo bagenu, ez genuke aurkituko koloreko populazioa eta kriminaltasun-tasaren arteko koerlazio esanguratsurik. "Beharbada" esaten dugu, lortutako emaitzak ikertutako kolektibora mugatzen baitira, edonolako estrapolaziorik egin gabe.

2 oharra. Notazioak azaltzen ez duen arren, kontutan hartu behar da **L** edo **R** matrizeek, ikertzen den kasuari egokiak zaizkion dimentsio eta aldagaiak izan behar dituztela.

Horrela, formula berbera edozein aldagai kopururekin erabil dezakegu.

Adibidez, X_1-en X_4-ren arteko koerlazio partziala, X_2-ren eragina kenduta kalkulatu nahi badugu, X_1, X_2 eta X_4 aldagaientzat definitutako **R** matrizea kontsidera dezagun. Horrela:

$$\mathbf{R} = \mathbf{R}(X_1, X_2, X_4) = \begin{pmatrix} 1 & 0.51 & 0.36 \\ & 1 & 0.60 \\ & & 1 \end{pmatrix}$$

Eta aldagaien ordenaketa berria kontutan izanik, kalkulatu nahi dugun koerlazioa izango da:

$$r_{13\bullet 2} = -\frac{\mathbf{R}_{13}}{\mathbf{R}_{11}^{1/2}\mathbf{R}_{33}^{1/2}} = 0.078$$

koerlazio-matrize berrian laugarren aldagaia hirugarrena izango baita.

Ikus daitekeenez, formula matrizialak, errepikapen formularekin batera etorri behar du; frogatzeko, adjuntoak garatu besterik ez da egin behar hiru aldagaien kasuan erraz frogatu ahal izateko. Kasu orokorra frogatu gabe geratzen da, kurtsoko asmoetatik kanpo baitago.

V.5.3. Koerlazio partzial eta erregresio-koefizienteen arteko erlazioa

Definizioan, $r_{12\bullet 3\cdots n}$-ren zeinuak X_1 eta X_2-ren arteko kobariantza (zuzena edo alderantzizkoa) adierazten du, beste n-2 aldagaien aldaketek eraginik ez dutenean; hau da, hauek konstanteak mantentzen direnean. Analitikoki gainera, zeinu hori b_{12} eta β_{12} koefizienteen X_1-en beste guztiekiko erregresio-koefizienteen zeinuen berdina izango du, eta beraz, interpretazio berbera, zeinu hori $\mathbf{L}_{12}$ adjuntuak mugatzen baitu.

Azkenik, (5.43), (5.44) eta (5.57) adierazpenetatik hurrengo erlazioak ondorioztatzen dira:

$$b_{12}b_{21} = \beta_{12}\beta_{21} = r^2_{12\bullet 3\cdots n} \qquad (5.58)$$

non b_{21} eta β_{21} bigarren aldagaiaren besteekiko erregresioaren koefizienteak diren.

V.A. Eranskina. ALDAGAI ESTATISTIKO n-KOITZEN MATRIZE-ESTATISTIKOAK

V.A.1. DATU-MATRIZEAK
V.A.2. BATEZBESTEKO-BEKTOREA: PROPIETATEAK
V.A.3. KOBARIANTZA MATRIZEA
V.A.3.1 Kobariantza matrizearen propietateak
V.A.4. KOERLAZIO-MATRIZEA
V.A.4.1. Koerlazio-matrizearen propietateak
V.A.5. DERIBAZIO BEKTORIALA

V.A.1. DATU-MATRIZEAK

Azkeneko ikasgaietan bi aldagaien arteko erlazioak aztertzeko tresna batzuk eman ditugu, baina hemendik aurrera gure ikasketako objektuak, hiru, lau,..... n aldagai izango dira eta ikasketak matrizialki egitea komenigarria da.

Has gaitezen, bada, matrizeak definitzen: X_1, X_2, ... X_n n aldagai kontsideratuz $\boldsymbol{X} = [X_1, X_2 \ldots X_n]$ beraiek osatzen duten zutabe-bektorea da:

Hots:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Aldagai horiek m indibiduoen gain neurketa konkretu batzuk edo balio ohartu batzuk hartzen dute, hauexek **X** datu-matrizea osatzen dutelarik.

Hau da:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_{m11} & x_{m2} & \ldots & x_{mn} \end{bmatrix} = \left[x_{ij} \right]$$

Datu-matrizeak, **X** alegia, eta aldagai bikoitz baten maiztasun-banaketaren matrizeak ez dute zerikusirik, biak matrizeak direla baizik; lehenengo kasuan, x_{ij}, elementua i. indibiduoari dagokion j. aldagaiarekiko balioa da eta bigarrenean, ordea, n_{ij} (x_i, y_j) balio bikotearen maiztasun absolutua.

j-garren aldagaiaren batezbesteko aritmetikoa eta desbidazio tipikoa

$$\bar{x}_j = \frac{\sum_{i=1}^{m} x_{ij}}{m} \quad \text{eta} \quad S_j = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2} \quad \text{izanik,}$$

Zentratze eta berreskalatze eragiketak

m x n ordenako **X** datu-matrizea maiz aldatzen da, zentratze, berreskalatze, tipifikatze edo bere zutabeen (aldagaien) normetan tipifikatze eragiketen bitartez.

Datu-matrize **zentratua, X** izendatuko duguna, horrela definitzen:

$$\mathbf{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \ldots\ldots & x_{1n} - \bar{x}_n \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \ldots\ldots & x_{2n} - \bar{x}_n \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} - \bar{x}_1 & x_{m2} - \bar{x}_2 & \ldots\ldots & x_{mn} - \bar{x}_n \end{bmatrix} = \left[x_{ij} - \bar{x}_j \right]_{i=1,\ldots,m}^{j=1,\ldots,n}$$

Datu-matrizea $\frac{1}{\sqrt{m}}$ -gatik biderkatuz, **X** izendatuko duguna, **zentratua eta berreskalatua** izango da:

$$\mathbf{X} = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{\sqrt{m}} & \frac{x_{12} - \bar{x}_2}{\sqrt{m}} & \ldots & \frac{x_{1n} - \bar{x}_n}{\sqrt{m}} \\ \frac{x_{21} - \bar{x}_1}{\sqrt{m}} & \frac{x_{22} - \bar{x}_2}{\sqrt{m}} & \ldots & \frac{x_{2n} - \bar{x}_n}{\sqrt{m}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{x_{m1} - \bar{x}_1}{\sqrt{m}} & \frac{x_{m2} - \bar{x}_2}{\sqrt{m}} & \ldots & \frac{x_{mn} - \bar{x}_n}{\sqrt{m}} \end{bmatrix} = \left[\frac{x_{ij} - \bar{x}_j}{\sqrt{m}} \right]_{i=1,\ldots,m}^{j=1,\ldots,n}$$

Zentratze eragiketak ez du datuen sakabanatze neurri sinple edo konposatuetan eragiten, lekualdatze bat besterik ez baita. Horregatik, bi matrizeentzat notazio berbera erabiliko dugu. Hala ere, zentratutako eta $\sqrt{m}$ -z berreskalatutako datu-matrizea, batzuetan erabiltzen da eta berreskalatzeak eragina izango du sakabanatzearen estatistikoetan.

Oharra: **X** deitura mantentzea notazioan gehiegikeria da, baina ez du eragozpenik, kasu bakoitzean bere gai orokorra zehaztuko baita.

V.A.2. BATEZBESTEKO-BEKTOREA: PROPIETATEAK

$\bar{\mathbf{x}}$-ikurraz adieraziko dugu eta aldagaien batezbestekoen zutabe-bektorea da.

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{bmatrix} = \left[\bar{x}_j \right]$$

Aldagai bakun baten batezbestekoaren linealtasun-propietateak, batezbesteko bektorerako zabaltzen dira berehala.

1. Propietatea

$$\left.\begin{aligned} \boldsymbol{Y} &= \mathbf{A}\boldsymbol{X} + \mathbf{b} \\ \mathbf{Y} &= \mathbf{X}\,\mathbf{A}^{\mathbf{T}} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \mathbf{b}^{\mathbf{T}} \end{aligned}\right\} \qquad \bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{x}} + \mathbf{b}$$

non: $\mathbf{A}$ matrizea $\mathbb{R}^n \rightarrow \mathbb{R}^n$ aplikazio lineal baten matrizea eta $\mathbf{b}$ lekualdatze-edo traslazio-bektorea diren.

Honen ondorioz zera esan dezakegu:
Aldagaien jatorri-aldatzeak eta unitate- edo eskala-aldatzeak batezbestekoetan daukaten eragina, zera da: beraiek ere linealki aldatzea. Hau da, batezbestekoetan transformazio lineala daukagu.

$$\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y} \Rightarrow \bar{z} = \bar{x} + \bar{y}$$

Aldagaien bektore bat beste bi aldagaien bektoreen batura denean baturaren batezbesteko-bektorea bi batugaien batezbesteko-bektoreen batura da.

V.A.3. KOBARIANTZA MATRIZEA

$(X_1, X_2 \ldots X_n)$ aldagai n-koitz baten bigarren ordenako momentu zentratuak bi eratakoak izan daitezke.

Aldagai bakun baten bazter-momentuak, adibidez:

$$m_{002.-0} = \frac{1}{m} \sum_{i=1}^{m} (x_{i3} - \bar{x}_3)^2 = S^2_{x_3}$$

Aldagai bikoitz baten bazter-momentuak, adibidez:

$$m_{101.-0} = \frac{1}{m} \sum_{i=1}^{m} (x_{i1} - \overline{x}_1)(x_{i3} - \overline{x}_3) = S_{x_1 x_3}$$

Hots bariantzak edo kobariantzak dira.

Aldagai baten bariantza aldagaiak berekiko duen kobariantza bezala uler dezakegu, eta definizioz $S_{ij} = S_{ji}$ da.

Hau guztia kontutan hartuz aldagai n-koitzaren 2. ordenako momentu zentratu guztiek matrize simetriko bat osatzen dute, kobariantza matrizea hain zuzen.

L ikurraz sinbolikoki adierazten dugu eta matrizialki honela lortzen da:

$$\mathbf{L} = \begin{vmatrix} S_1^2 & S_{12} & \dots & S_{1n} \\ S_{12} & S_2^2 & \dots & S_{2n} \\ \dots & \dots & \dots & \dots \\ S_{1n} & S_{2n} & \dots & S_n^2 \end{vmatrix} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1n} \\ l_{12} & l_{22} & \dots & l_{2n} \\ \dots & \dots & \dots & \dots \\ l_{1n} & l_{2n} & \dots & l_{mm} \end{bmatrix}$$

$X = [x_{ij} - \overline{x}_j]_{i,j}$, datu-matrize zentratua izanik, X_1, ..., X_n aldagaien hurrengo bariantza eta kobariantza matrizea lortzen da:

$$\mathbf{L} = \frac{1}{m} X^T X$$

$\mathbf{X} = \left[\frac{x_{ij} - \overline{x}_j}{\sqrt{m}} \right]_{i,j}$, zentratu eta berreskalatutako datu-matrizea izanik, L horrela lortzen da:

$$\mathbf{L} = \mathbf{X^T X}$$

V.A.3.1. Kobariantza matrizearen propietateak

1. propietatea

$$\left.\begin{array}{l} \boldsymbol{Y = A\,X + b} \\ \mathbf{Y = X\,A^T + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} b^T} \end{array}\right\} \quad \Rightarrow \quad \mathbf{L(Y) = A\ L\,(X)\,A^T}$$

Ikusten dugunez eskala aldatzeak badauka eragina kobariantza matrizean, baina ez jatorri aldatzeak.

Oharra: gogora dezagun $Y = aX + b$ izanik, bariantzen arteko erlazioa zera zela:

$$S_y^2 = a^2 S_x^2$$

2. propietatea:

Kobariantza matrizeak definitu positiboak edo erdidefinitu positiboak dira.

Matrize-algebran honek suposatzen du azpimatrize diagonal guztiak definitu positiboak edo erdidefinitu positiboak direla.

Datu-matrize tipifikatua

Tipifikatzeak edo norman tipifikatzeak aurretiko zentratzeaz gain, aldagaietan eskala aldatzea suposatzen du. Lehenengo kasuan, transformazioa egin ondoren, aldagaiek 1 baliodun desbidazio tipikoa dute, eta bigarrenean, m dimentsioko bektoreek osatzen dituzte, norma edo moduloa 1 dutelarik.

Horrela hauxe daukagu:

$$\mathbf{T} = \begin{bmatrix} \dfrac{x_{11} - \bar{x}_1}{s_1} & \dfrac{x_{12} - \bar{x}_2}{s_2} & \cdots & \dfrac{x_{1n} - \bar{x}_n}{s_n} \\ \dfrac{x_{21} - \bar{x}_1}{s_1} & \dfrac{x_{22} - \bar{x}_2}{s_2} & \cdots & \dfrac{x_{2n} - \bar{x}_n}{s_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{x_{m1} - \bar{x}_1}{s_1} & \dfrac{x_{m2} - \bar{x}_2}{s_2} & \cdots & \dfrac{x_{mn} - \bar{x}_n}{s_n} \end{bmatrix} = \left[\frac{x_{ij} - \bar{x}_j}{S_j} \right]_{i=1,\ldots,m}^{j=1,\ldots,n} = \left[t_{ij} \right]_{i,j}$$

non $S_{tj}^2 = S_{tj} = 1$ den, edota:

$$\mathbf{N} = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{s_1\sqrt{m}} & \frac{x_{12} - \bar{x}_2}{s_2\sqrt{m}} & \cdots & \frac{x_{1n} - \bar{x}_n}{s_n\sqrt{m}} \\ \frac{x_{21} - \bar{x}_1}{s_1\sqrt{m}} & \frac{x_{22} - \bar{x}_2}{s_2\sqrt{m}} & \cdots & \frac{x_{2n} - \bar{x}_n}{s_n\sqrt{m}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{x_{m1} - \bar{x}_1}{s_1\sqrt{m}} & \frac{x_{m2} - \bar{x}_2}{s_2\sqrt{m}} & \cdots & \frac{x_{mn} - \bar{x}_n}{s_n\sqrt{m}} \end{bmatrix} = \left[\frac{x_{ij} - \bar{x}_j}{S_j\sqrt{m}} \right]_{i=1,\ldots,m}^{j=1,\ldots,n} = \left[n_{ij} \right]_{i,j}$$

non $\left\| \mathbf{n}_j \right\| = 1$.

V.A.4. KOERLAZIO-MATRIZEA

Aldagai tipifikatuen arteko kobariantza matrizea da.
Hau da:

$$\mathbf{L}(\boldsymbol{T}) = \frac{1}{m}\mathbf{T}^{\mathrm{T}}\mathbf{T} = \frac{1}{m}\left[\frac{x_{ij} - \bar{x}}{S_j} \right]^{\mathrm{T}} \left[\frac{x_{ij} - \bar{x}_j}{S_j} \right] = \mathbf{R}$$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1n} \\ r_{12} & 1 & \ldots & r_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ r_{1n} & r_{2n} & \ldots & 1 \end{bmatrix}$$

halaber: $\mathbf{R} = \mathbf{N}^{\mathbf{T}}\mathbf{N}$ $\qquad$ $\mathbf{N} = \left[\frac{x_{ij} - \bar{x}j}{S_j\sqrt{m}} \right]$

normatutako datu-matrizea izanik.

X datu-matrizearen eta **T** datu-matrize tipifikaturen arteko erlazioa hauxe da:

$$\mathbf{X} = \mathbf{T}\,\mathbf{D}_{\mathbf{S}}^{\mathbf{T}} + \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \bar{\mathbf{x}}^{\mathrm{T}}$$

Halaber, $\boldsymbol{X}$ aldagaien bektorea eta $\boldsymbol{T}$ aldagai tipifikatuen artekoa:

$$\boldsymbol{X} = \mathbf{D}_S \; \boldsymbol{T} + \bar{\mathbf{x}}$$

Espresio honek adierazten duena zera da: transformazio lineal diagonal bat dugula ($\mathbb{R}^n$ espazio bektorialean definitua), transformazio hau, $\mathbf{D}_s$ matrize diagonalak definitu baitu.

$$\mathbf{D}_S = \begin{bmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & S_n \end{bmatrix}$$

Dakusagun $\mathbf{L}(\boldsymbol{X})$ eta $\mathbf{L}(\boldsymbol{T})$ kobariantza matrizeen arteko erlazioa:

$$\mathbf{L}(\boldsymbol{X}) = \mathbf{D}_S \; \mathbf{L}(\boldsymbol{T}) \; \mathbf{D}_S^t$$

Hots:

$$\mathbf{L} = \mathbf{D}_S \; \mathbf{R} \; \mathbf{D}_S^t$$

Kobariantza eta koerlazio-matrizeen arteko erlazioa.

Estentsiboki:

$$\mathbf{L} = \begin{bmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & S_n \end{bmatrix} \begin{bmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{12} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{1n} & r_{2n} & \dots & 1 \end{bmatrix} \begin{bmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & S_n \end{bmatrix}^T$$

Dakusagun azkenik matrize hauen determinanteen eta adjuntuen arteko erlazioak.

$\mathbf{L}$ eta $\mathbf{R}$ matrizeen arteko elementuetan $S_{ij} = S_i \, S_j \, r_{ij}$ erlazioa betetzen dela ikusirik, ondoko erlazioak, erraz ateratzen ditugu:

$\|\mathbf{L}\| = S_1^2 S_2^2 \ldots S_n^2 \|\mathbf{R}\|$	$\|\mathbf{L}\|$ eta $\|\mathbf{R}\|$, $\mathbf{L}$ eta $\mathbf{R}$ matrizeen determinanteak izanik
$\mathbf{L}_{11} = S_2^2 S_3^2 \ldots S_n^2 \mathbf{R}_{11}$	$\mathbf{L}_{11} = l_{11}$ elementuaren adjuntua $\mathbf{R}_{11} = r_{11}$ elementuaren adjuntua
$\mathbf{L}_{12} = S_1 S_2 S_3^2 \ldots S_n^2 \mathbf{R}_{12}$	$\mathbf{L}_{12} = l_{12}$ elementuaren adjuntua $\mathbf{R}_{12} = r_{12}$ elementuaren adjuntua

V.A.4.1. Koerlazio-matrizearen propietateak

Gogoratu behar dugu (II.3.2.) koerlazioak, aldagai tipifikatuen arteko kobariantzak direla.

Honela, koerlazio-matrizea kobariantza matrize mota bat izatean, kobariantza matrizeen propietateak ditu.

Bestalde, koerlazio-matrizea, bi aldagaien arteko koerlazioa bezala (II.3) inbariantea da edozein transformazio linealen aurrean.

V.A.5. DERIBAZIO BEKTORIALA

Ondoren, eskema moduan, deribazio bektorialeko hurrengo formulak gogoratuko ditugu, R-ko beste hainbat funtziorekin erlazionatuz eta horrela bere ikasketa erraztuz.

Karratua: $\frac{d(x^2)}{dx} = 2x$	$\frac{\partial(\mathbf{x}^T\mathbf{x})}{\partial\mathbf{x}} = 2\mathbf{x}$
Karratua: $\frac{\partial(xy)}{\partial x} = \frac{\partial(yx)}{\partial x} = y$	$\frac{\partial(\mathbf{y}^T\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{y})}{\partial\mathbf{x}} = \mathbf{y}$
Biderkadura eskalar batez: $\frac{d(ax)}{dx} = a$	$\frac{\partial(\mathbf{Ax})}{\partial\mathbf{x}} = \mathbf{A}^T$
Forma koadratikoa: $\frac{d(ax^2)}{dx} = 2ax$	$\frac{\partial(\mathbf{x}^T\mathbf{Ax})}{\partial\mathbf{x}} = 2\mathbf{Ax}$, (asimetrikoa)
Eskalarra: $\frac{d(kf(x))}{dx} = k\frac{df(x)}{dx}$	$\frac{\partial(kf(\mathbf{x}))}{\partial\mathbf{x}} = k\frac{\partial f(\mathbf{x})}{\partial\mathbf{x}}$

Deribazio matrizialaren formulen ikerketa sakonagoa egiteko, ikus adibidez, Martin Pliegoren Introducción a la Estadística Económica y Empresarial, (1994).

VI. ZENBAKI-INDIZEAK

VI.1. INDIZE SINPLEAK
VI.2. PONDERAZIO GABEKO INDIZE KONPLEXUAK
VI.2.1. Batezbesteko aritmetiko sinplearen metodoa
VI.2.2. Batezbesteko agregatu sinplearen metodoa
VI.3. INDIZE KONPLEXU PONDERATUAK
VI.3.1. Balioen, prezioen eta kopuruen indizeak
VI.3.1.1. LASPEYRES-en indizeak
VI.3.1.2. PAASCHE-en indizeak
VI.3.1.3. FISHER-en indizeak
VI.3.1.4. Propietate eta erlazio batzuk: bateragarritasuna eta alderantzizkotasuna
VI.3.1.5. Kalkulua
VI.4. INDIZE KONPLEXUEN ERAIKETAN SORTZEN DEN ZENBAIT ERAGOZPEN
VI.4.1. Aldagaien hautapena
VI.4.2. Somatutako leku eta denboraren hautapena
VI.4.3. Talde eta azpitaldeen hautapena
VI.4.4. Oinarri-denboraren hautapena
VI.4.5. Formula eta ponderazioen hautapena
VI.4.6. Indizearen esangura eta zabaldura
VI.5. ERAGOZPEN BEREZI BATZUK
VI.5.1. Oinarri-aldaketa indize sinpleetan
VI.5.2. Berriztapen eta loturak indize konplexuetan
VI.6. ZENBAKI-INDIZEEN APLIKAPENAK
VI.6.1. Kontsumo-prezioen indizeak
VI.6.2. Moneta-unitate arruntetan dauden magnitudeen deflazioa

VI.1. INDIZE SINPLEAK

Aldagai baten gora-beherakadak aztertzeko, **indize sinpleak** erabiltzen dira, eta aldagai batzuen gora-beherakadak "batera" aztertzeko **indize konplexuak**.

Hemen denboran zeharreko gora-beherakadak aztertuko dira. Erreferentzia gisa hartzen dugun denborari, **oinarria** deritzogu.

Erreferentziaren balioarekiko aldagaiaren balio bakoitzaren portzentaiak baino ez dira indize sinpleak.

Portzentaia hauen bidez, neurtzeko unitatea desagertzen da eta aldagai baten gora-beherakadak autonomoki azter daitezke. Honela indizeak aldagai-segiden gora-beherakadak errazten ditu (originalki unitate desberdinetakoak).

Izan bedi $x_0, x_1 \ldots x_t$ aldagaiaren denborazko balio-segida bat; hots, 0, 1, ... t denboretan neurtutako balioak.

$t = 0$ oinarritzat harturik eta x_0 balio erreferentziala, indize sinpleak ondoko taulan agertzen diren bezala kalkulatzen dira:

Denbora: t	Balio-segida: x_t	Indize Sinpleak
0	x_0	$I_0 = 100$
1	x_1	$I_1 = \frac{x_1}{x_0} \cdot 100$
2	x_2	$I_2 = \frac{x_2}{x_0} \cdot 100$
.	.	.
.	.	.
.	.	.
t	x_t	$I_t = \frac{x_t}{x_0} \cdot 100$

x_0 balio erreferentzial gisa “urte normal” bati dagokiona hartzea, oso garrantzitsua da.

Adibidez: Instintiboki uzta-bilketa bat ona edo txarra dela esatean, guretzat uzta-bilketa “normalak” direnekin konparaturik egiten dugu.

VI.2. PONDERAZIO GABEKO INDIZE KONPLEXUAK

n balioren aldakuntzak batera aztertu nahi ditugunean, indize konplexuak eraiki behar ditugu. Horiek n balio-segidak, balio-segida bakar batera laburtzen dituzte, non balioak, oinarritzat hartzen den urte batera erreferiturik dauden eta “konplexu” guztiaren mugimendua ikusarazten diguten.

Goazen, bada, ponderazio gabeko indize konplexuak kalkulatzeko bi metodo ikustera.

VI.2.1. Batezbesteko aritmetiko sinplearen metodoa

Indize konplexu hau, denbora bakoitzari dagozkion **indize sinpleen** batezbesteko aritmetikoa da.

Hots: t denboran $x_1, x_2, ... x_n$ aldagaiei dagozkien $x_{1t}, x_{2t}, ... x_{it}, ... x_{nt}$ n balio-segidak baldin badira, non $t = 0,1,2,...m$ den I_{1t} I_{2t} ..... I_{nt} indize sinpleak kalkulatu ondoren, indize konplexua formula honetaz kalkulatzen den:

$$S_t = \frac{\sum_{i=1}^{n} I_{it}}{n} = \frac{\sum_{i=1}^{n} \frac{x_{it}}{x_{i0}} 100}{n}$$

Momentu batean, indize sinpleen elkarrekiko garrantzia ez hartzeagatik, ponderazio gabeko indizea da, orduan, **zenbat eta aldagai gehiagok parte hartu, orduan eta hobea gertatzen da metodo hau**.

Estatistika institutuan, urtez urteko aldagai kopuru handi batez Estatu espainolerako metodo honetaz edo metodo ponderatuen bidez kalkulatzen diren indizeak, ez dira hain desberdinak.

VI.2.2. Batezbesteko agregatu sinplearen metodoa

Metodo honen bidez, une bakoitzerako aldagai guztien balioen batura egiten da eta lortzen den segidan indize sinpleak bakarrik kalkulatzen dira.

Formula honetaz kalkulatzen dira:

$$B_t = \frac{\sum_{i=1}^{n} x_{it}}{\sum_{i=1}^{n} x_{i0}} = 100$$

Metodo hau, **ezin erabil daiteke aldagaiak unitate desberdinetan neurtuak direnean**, ez baitu zentzu askorik unitate desberdinetan neurtuak izan diren balioak batzea.

Ponderazio gabeko indize hau aurkez daiteke indize sinpleen indize ponderatu bezala ere, ponderazioak oinarri denborako balioak direlarik.
(Ikus batezbesteko aritmetiko ponderatua 27. orrialdean)

$$B_t = \frac{\sum_{i=1}^{n} I_{it} x_{i0}}{\sum_{i=1}^{n} x_{i0}}$$

Ondoko taulan ponderazio gabeko indize konplexu hauen formulak dauzkagu hedaturik.

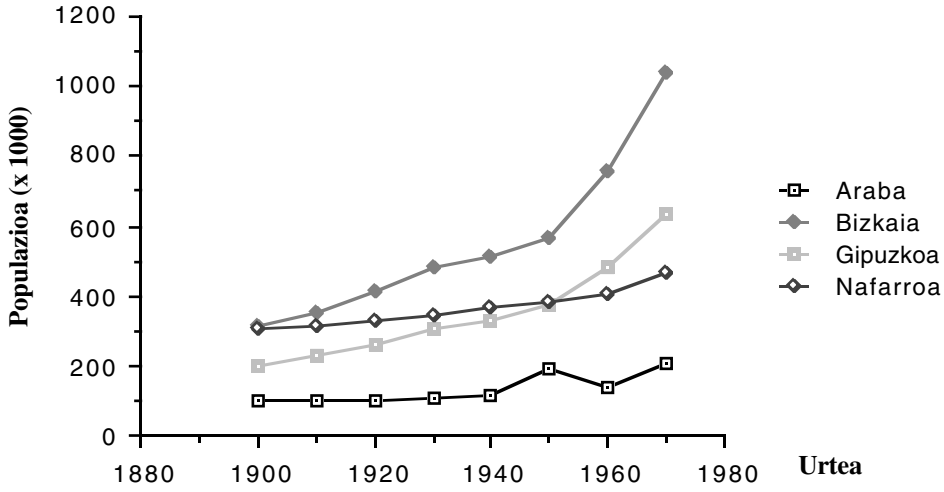
x / t	$x_1, x_2, \dots x_i, \dots x_n$	I_i (Indize sinpleak)	S_t	$\sum_i x_{it} B_t$
0	$x_{10}\ x_{20} \dots x_{i0} \dots x_{n0}$	$I_{10}\ I_{20} \dots I_{n0}$	$S_0 = \frac{\sum_i I_{i0}}{n}$	$\sum_i x_{i0} B_0 = \frac{\sum_i x_{i0}}{\sum_i x_{i0}} 100$
1	$x_{11}\ x_{21} \dots x_{i1} \dots x_{n1}$	$I_{11}\ I_{21} \dots I_{n1}$	$S_1 = \frac{\sum_i I_{i1}}{n}$	$\sum_i x_{i1} B_1 = \frac{\sum_i x_{i1}}{\sum_i x_{i1}} 100$
. . .			. . .	. . .
t	$x_{1t}\ x_{2t} \dots x_{it} \dots x_{nt}$	$I_{1t}\ I_{2t} \dots I_{nt}$	$S_t = \frac{\sum_i I_{it}}{n}$	$\sum_i x_{it} B_t = \frac{\sum_i x_{it}}{\sum_i x_{it}} 100$
. .				. .
m	$x_{1m}\ x_{2m} \dots x_{im} \dots x_{nm}$	$I_{1m}\ I_{2m} \dots I_{nm}$	S_m	B_m

Adibidea: Hego Euskal Herriko populazioaren datu hauen bidez, populazio igoerak ikusiko ditugu herrialde bakoitzean.

Datuak:

Urtea	Araba	Bizkaia	Gipuzkoa	Nafarroa
t	x_1	x_2	x_3	x_4
1.900	96.385	311.361	195.850	307.669
1.910	97.181	349.932	226.684	312.235
1.920	98.668	409.550	258.557	329.875
1.930	104.176	485.205	302.329	345.883
(1) 1.940	112.876	511.135	331.753	369.618
1.950	188.012	569.188	374.040	382.932
1.960	138.934	754.383	478.337	402.042
1.970	204.323	1.043.311	631.003	464.867
1.980	258.200	1.179.105	692.586	507.367
1.990	272.101	1.153.622	675.529	527.318

1. Taula



1. Grafikoa

1. GRAFIKOAN IKUSTEN DIREN GAUZA NABARMENENAK

Grafiko honek, herrialde bakoitzeko **populazioaren hazkunde absolutua** adierazten digu.

Ondorio hauek atera genitzake:

1.900. urtean Nafarroako eta Bizkaiko populazioak berdintsuak ziren, Nafarroak XIX. mendean zuen populazioa askoz handiagoa izan arren.

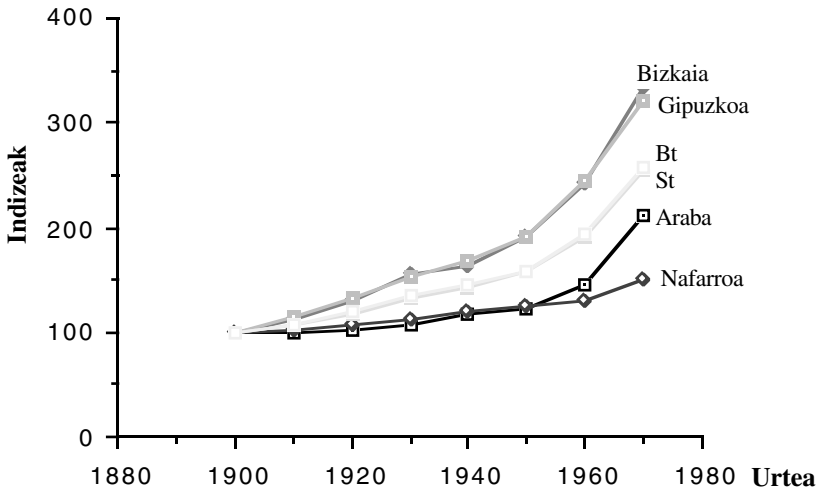
Hamarkada bakoitzean, ibilbidearen malda, gehitzen den biztanle-kopuruaren araberakoa da, (non gehikuntza hamarkadaren hasierako populazio absolutuarekiko baita).

Ikusten denez, Nafarroa eta Arabaren ibilbideak antzekoak dira alde batetik, eta bestetik, Bizkaia eta Gipuzkoarenak ere oso antzekoak dira.

Indizeak:

	Urteak	I_{1t}	I_{2t}	I_{3t}	I_{4t}	S_t	$\sum_i x_{it}$	B_t
(2)	1.900	100'00	100'00	100'00	100'00	100'00	911'265	100'00
	1.910	100'83	112'38	115'74	101'48	107'61	986'032	108'20
	1.920	102'37	131'53	132'02	107'22	118'28	1.086'670	119'24
	1.930	108'08	155'83	154'37	112'42	132'67	1.237'593	135'81
	1.940	117'11	164'16	169'39	120'13	142'70	1.325'364	145'44
	1.950	122'44	191'48	190'98	124'46	157'34	1.444'172	158'48
	1.960	144'15	242'28	244'23	130'67	190'33	1.773'696	194'64
	1.970	211'98	335'08	322'18	151'09	255'08	2.343'504	257'17

2. Taula



2. Grafikoa

2. GRAFIKOAN IKUSTEN DIREN GAUZA NABARMENENAK

Alde batetik grafiko honetan, herrialde bakoitzari dagokion populazio-indize sinpleen ibilbidea daukagu.

Indize sinple horien ibilbideek, **populazioaren gehikuntza erlatiboa, 1.900 urteko populazioarekiko**, adierazten digute, herrialde guztietako populazioa 1.900 urtean, 100 dela suposatuz.

1.900 urtetik zenbat eta gehiago urruntzen garen, orduan eta ibilbide horien maldek esangura gutxiago izango dute.

Beste batetik, Hego Euskal Herriko populazio-indize konplexu bien (S_t, B_t) ibilbideak ditugu (erdikoak).

Azkenik, hamarkada bakoitzeko urtez urteko hazkunde-tasak bilatu eta 3. grafikoan adierazten dira grafikoki.

Hots:

$$i_{t,t-1} = \frac{x_{i,t}}{x_{i,t-1}} = (1 + \alpha)$$

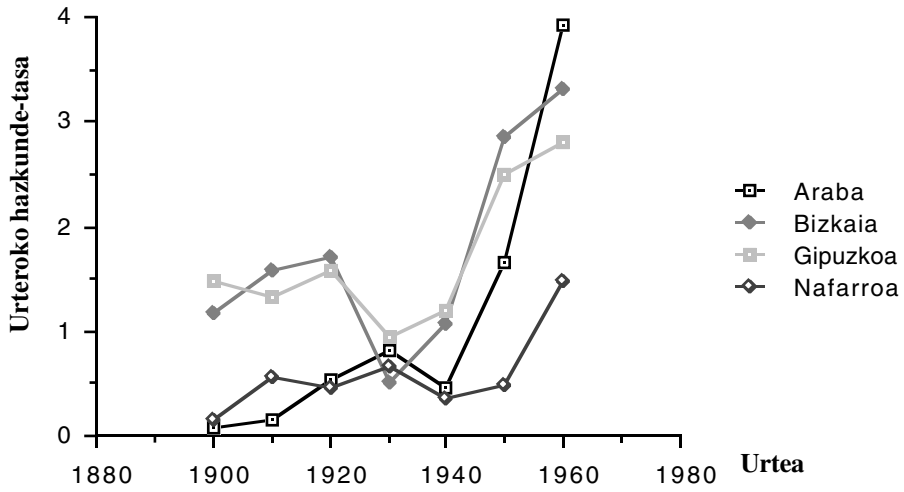
Hemen $i_{t,t-1}$ delakoa momentu baten aurrekoarekiko indize sinplea da. Hau da, t-1 uneko unitate bat, $1 + \alpha$ bihurtu da t unean, orduan hamarkada batean 0 uneko unitate bat, $(1 + \alpha)^{10}$ bihurtu da 10 urtetan, eta exenplu honetan datuak hamarkadaz ditugunez, **urteroko hazkunde-tasa** ondoko formulen bidez kalkulatzen dugu:

$$\alpha + 1 = \sqrt[10]{\frac{x_{i,t}}{x_{i,t-10}}} \quad \text{eta} \quad \alpha = \sqrt[10]{\frac{x_{i,t}}{x_{i,t-10}}} - 1$$

Ondoko taulan, urteroko hazkunde-tasak adierazten dira:

α x 100	1.900/ 1.910	1.910/ 1.920	1.920/ 1.930	1.930/ 1.940	1.940/ 1.950	1.950/ 1.960	1.960/ 1.970
ARABA	0'08	0'15	0'54	0'81	0'45	1'65	3'93
BIZKAIA	1'17	1'59	1'71	0'52	1'08	2'86	3'30
GIPUZKOA	1'47	1'32	1'58	0'93	1'21	2'49	2'80
NAFARROA	0'15	0'55	0'47	0'67	0'35	0'49	1'47

3.Taula.



3. Grafikoa

3. GRAFIKOAN IKUSTEN DIREN GAUZA NABARMENENAK

Grafiko honetan, herrialde bakoitzari dagokion hazkunde-tasen ibilbidea daukagu.

Hazkunde-tasen ibilbideek **populazioaren gehikuntza erlatiboa adierazten** digute, kasu honetan gehikuntza **hamarkada bakoitzarekikoa** delarik.

Globalki 7 hamarkada horietan, Bizkaia eta Gipuzkoari dagozkien hazkunde--tasen ibilbideek dute azelerazio gehien.

Geldiera handia nabaritzen da lau herrialdeetan, 1.936 urte inguruan (gerra izan liteke arazo honen zergatia).

Lehenengo hamarkadan nahiko geldikorrak dira hazkunde-tasak bai Araban, bai Nafarroan; gero bietan hazkunde handia nabaritzen da, ordea, 1.930-1.940 hamarkadan, justu Bizkaia eta Gipuzkoan geldiera handia denean (migrazioa eta gerra izan litezke arazo honen zergatia?).

Azken hamarkadetan izugarrizko azelerazioa nabaritzen da lau herrialdeetan; kezkagarriak Bizkaia eta Gipuzkoaren kasuak, azkenean Arabak, Bizkaia eta Gipuzkoaren saturazioa jasotzen duelarik.

Ariketa: Azken bi hamarkadako populazioaren datuak (1. taula) kontutan hartuz, lor itzazu dagozkien indizeak, hazkunde-tasak, eta adieraz itzazu 3 grafikoetan.

VI.3. INDIZE KONPLEXU PONDERATUAK

Orain arte, ponderazio gabeko indizeak aztertu ditugu. Baina, adibidez, arroza garia baino bi aldiz gehiago kontsumitzen bada, prezioen indize baten kalkuluan garrantzi handiagoa eman behar zaie arrozaren prezioaren aldaketei gariarenei baino.

Hau ponderaketaren bidez lortzen da.

Prezioen indize bat kalkulatzean, ondasun bakoitza, kontsumitzen den kopuruaz ponderatzen bada, pauta batzuk jarraituz egingo dugu, eta artikulu edo ondasun batentzat aukeratuak izan diren pautak besterentzat ere errespetatu behar dira ahal bada.

Adibidez, hileroko janari prezioen indize batean, esnearen prezioa bere kantitateaz ponderatzen bada eta kopuru hau famili bati dagokiona baldin bada, (senitarte kopuru mugatu eta errenta-tarte baten barnekoa), orduan, ogiari dagokion ponderazioak ere irizpide berdinetan oinarritua izan behar du.

VI.3.1. Balioen prezioen eta kopuruen indizeak

"p" delako ondasun baten prezioa bada eta "k" beraren kopurua (saldua, produzitua...) orduan, "b" delako ondasunaren balioa dela esango dugu eta honela kalkulatuko dugu:

$$b = p \cdot k$$

0,1,.....t,....,m denboretan, ondasun baten prezioak
$p_0, p_1, ...p_t, ...p_m$ eta dagozkien kopuruak
$k_0, k_1, ...k_t, ...k_m$ baldin badira, orduan "denborazko balio-segida" hau izango da:

$$\underline{\quad b_t \quad}$$

$b_0 = p_0k_0$

$b_1 = p_1k_1$

.
.
.

$b_t = p_tk_t$

.
.
.

$b_m = p_mk_m$

Baina denbora iragatean k = kte bada, b_t denborazko balio-segidak, prezioaren gora-beherakadak adieraziko ditu bakarrik. Berdin, denbora iragatean p = kte bada, b_t balio-segidak kopuruaren gora-beherakadak adieraziko ditu bakarrik.

Orduan, ondasun bati dagozkion hiru "denborazko balio-segida" desberdin idatz ditzakegu.

b_t	b_t (k)	b_t (p)
p_0k_0	pk_0	p_0k
p_1k_1	pk_1	p_1k
.		
.		
p_tk_t	pk_t	p_tk
.		
.		
p_mk_m	pk_m	p_mk

"O" ondasun bati dagozkion (balio, prezio eta kopuruen) segida horiek balio errealetan adieraziak direnez gero, beste "O_i" ondasun batzuei dagozkienak ere batu ditzakegula, argi ikusten da.

Suposa dezagun O_1 O_2 ..... O_i ..... O_n n ondasun desberdin ditugula, (non O_i ondasun bakoitzaren balio, prezio eta kopuruen denborazko balio-segidak ezagunak zaizkigun), n ondasun horiek, batera osatzen duten "konplexuaren indizeak" edo "indize konplexuak" kalkulatzerakoan, balio-segida desberdinen batuketak, laburki, honela adieraz daitezke:

$\sum_{i=1}^{n} b_{it}$	$\sum_{i=1}^{n} b_{it}(k)$	$\sum_{i=1}^{n} b_{it}(p)$
$\sum_{i} p_{i0}k_{i0}$	$\sum_{i} p_{i}k_{i0}$	$\sum_{i} p_{i0}k_{i}$
$\sum_{i} p_{i1}k_{i1}$	$\sum_{i} p_{i}k_{i1}$	$\sum_{i} p_{i1}k_{i}$
$\vdots$	$\vdots$	$\vdots$
$\sum_{i} p_{it}k_{it}$	$\sum_{i} p_{i}k_{it}$	$\sum_{i} p_{it}k_{i}$
$\vdots$	$\vdots$	$\vdots$
$\sum_{i} p_{im}k_{im}$	$\sum_{i} p_{i}k_{im}$	$\sum_{i} p_{im}k_{i}$

(Bi azkenetako segidetan, $p_i = kte, \forall_i$ eta $k_i = kte, \forall_i$ dira).

Beraz, **indize konplexuak** lortzeko konplexuari dagozkion (balio, prezio eta kopurua) hiru **denborazko balio segida** 1. horien **indize sinpleak kalkulatu** behar ditugu.

Hots oinarri-denbora t = 0 bada:

I_t^b	I_t^k	I_t^p
$\frac{\sum_{i} p_{i0}k_{io}}{\sum_{i} p_{i0}k_{io}} 100$	$\frac{\sum_{i} (p_i)k_{io}}{\sum_{i} (p_i)k_{io}} 100$	$\frac{\sum_{i} p_{i0}(k_i)}{\sum_{i} p_{i0}(k_i)} 100$
$\frac{\sum_{i} p_{i1}k_{i1}}{\sum_{i} p_{i0}k_{i0}} 100$	$\frac{\sum_{i} (p_i)k_{i1}}{\sum_{i} (p_i)k_{i0}} 100$	$\frac{\sum_{i} p_{i1}(k_i)}{\sum_{i} p_{i0}k_{i0}} 100$
$\vdots$	$\vdots$	$\vdots$

1. Lehenengoa prezio eta kopuruaren funtzioa da, bigarrena, kopuruaren funtzioa, eta hirugarrena prezioarena.

$$\frac{\sum_i p_{it}k_{it}}{\sum_i p_{i0}k_{i0}}100 \qquad \frac{\sum_i (p_i)k_{it}}{\sum_i (p_i)k_{i0}}100 \qquad \frac{\sum_i p_{it}(k_i)}{\sum_i p_{i0}(k_i)}100$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$\frac{\sum_i p_{im}k_{im}}{\sum_i p_{i0}k_{i0}}100 \qquad \frac{\sum_i (p_i)k_{im}}{\sum_i (p_i)k_{i0}}100 \qquad \frac{\sum_i p_{im}(k_i)}{\sum_i p_{i0}(k_i)}100$$

I_t^b I_t^k I_t^p aurreko indizeei, balio, prezio eta kopuruen **indize konplexu agregatuak** deritze, erabiltzen den agregazioa ponderatua bada ere.

Baina orain problema bat sortzen zaigu:

I_t^k "kopuruen indize konplexua" lortzeko, O_i ondasun bakoitzaren zein prezio hartuko dugu konstantetzat?

Edota:

I_t^p "prezioen indize konplexua" lortzeko, O_i ondasun bakoitzaren zein kopuru hartuko dugu konstantetzat?

Problema honetarako nahi beste ebazpide edo soluzio aurki daitezke, baina praktikan bi inposatu dira: LASPEYRES-ena eta PAASCHE-rena.

I_t^k eta I_t^p lortzeko, horiek urratutako bide desberdinak ikus ditzagun ondoren.

VI.3.1.1. LASPEYRES-en indizeak

I_t^k eta I_t^p prezio eta kopuruen indize konplexuak kalkulatzeko, O_i ondasun bakoitzaren kasuan konstantetzat hartzen ditugun kopuru eta prezioak, oinarri-denboran sortutakoak dira.

Hots, $k_i = k_{i0}$ eta $p_i = p_{i0}$

Honela lortutako indize konplexuei, **LASPEYRES-en prezio eta kopuruen indize konplexuak** deritzegu eta sinbolikoki L_t^p eta I_t^k adieraziko ditugu.

Hots:

$$\begin{array}{cc} L_t^p & L_t^k \\ \hline \\ \dfrac{\sum_i p_{i0}(k_{i0})}{\sum_i p_{i0}(k_{i0})}\ 100 & \dfrac{\sum_i (p_{i0})k_{i0}}{\sum_i (p_{i0})k_{i0}}\ 100 \\ \\ \dfrac{\sum_i p_{i1}(k_{i0})}{\sum_i p_{i0}(k_{i0})}\ 100 & \dfrac{\sum_i (p_{i0})k_{i1}}{\sum_i (p_{i0})k_{i0}}\ 100 \\ \vdots & \vdots \\ \dfrac{\sum_i p_{it}(k_{i0})}{\sum_i p_{i0}(k_{i0})}\ 100 & \dfrac{\sum_i (p_{i0})k_{it}}{\sum_i (p_{i0})k_{i0}}\ 100 \\ \vdots & \vdots \\ \dfrac{\sum_i p_{im}(k_{i0})}{\sum_i p_{i0}(k_{i0})}\ 100 & \dfrac{\sum_i (p_{i0})k_{im}}{\sum_i (p_{i0})k_{i0}}\ 100 \end{array} \qquad \textbf{(1)}$$

Oinarrizko eragiketa batez, erraz ikus daiteke (**1**) formulak, ondoren datozen (**2**) formula horien baliokideak direla:

$$\begin{array}{cc} \dfrac{L_t^p}{100} & \dfrac{L_t^k}{100} \\ \\ \dfrac{\sum_i \frac{p_{i1}}{p_{i0}}(p_{i0}\ k_{i0})}{\sum_i (p_{i0}\ k_{i0})}\ 100 & \dfrac{\sum_i \frac{k_{i1}}{k_{i0}}(p_{i0}\ k_{i0})}{\sum_i (p_{i0}\ k_{i0})}\ 100 \\ \vdots & \vdots \end{array}$$

$$\frac{\sum_i \frac{p_{it}}{p_{i0}}(p_{i0}\,k_{i0})}{\sum_i (p_{i0}\,k_{i0})}\,100 \qquad \frac{\sum_i \frac{k_{it}}{k_{i0}}(p_{i0}\,k_{i0})}{\sum_i (p_{i0}\,k_{i0})}\,100 \qquad \textbf{(2)}$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\sum_i \frac{p_{im}}{p_{i0}}(p_{i0}\,k_{i0})}{\sum_i (p_{i0}\,k_{i0})}\,100 \qquad \frac{\sum_i \frac{k_{im}}{k_{i0}}(p_{i0}\,k_{i0})}{\sum_i (p_{i0}\,k_{i0})}\,100$$

Ikusten denez, oinarri-denborako prezio edo kopuruaz biderkatzen eta zatitzen da zenbakitzailearen batugai bakoitzean.

Era honetara, batezbesteko aritmetiko ponderatu bezala, (**2**) formulak oso erabilgarriak zaizkigu: O_i ondasun bakoitzaren indize sinpleak erabiliz LASPEYRES-en prezio edo kopuruen indize konplexuak kalkulatzen ditugu.

VI.3.1.2. PAASCHE-ren indizeak

I_t^p eta I_t^k prezio eta kopuruen indize konplexuak kalkulatzeko, O_i ondasun bakoitzaren kasuan indizea kalkulatzen dugun denborako, kopurua eta prezioa hartzen ditugu konstantetzat.

Hots: $k_i = k_{it}$ eta $p_i = p_{it}$

Honela lortutako indize konplexuei, **PAASCHE-ren prezio eta kopuruen indize konplexuak** deritzegu eta sinbolikoki P_t^p eta P_t^k adieraziko ditugu.

Hots:

$$P_t^p \qquad\qquad P_t^k$$

$$\frac{\sum_i p_{i1}(k_{i1})}{\sum_i p_{i0}(k_{i1})}\,100 \qquad \frac{\sum_i (p_{i1})\,k_{i1}}{\sum_i (p_{i1})\,k_{i0}}\,100$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\sum_i p_{it}(k_{it})}{\sum_i p_{i0}(k_{it})}\ 100 \qquad \frac{\sum_i (p_{it})\ k_{it}}{\sum_i (p_{it})\ k_{i0}}\ 100 \qquad \textbf{(1)}$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\sum_i p_{im}(k_{im})}{\sum_i p_{i0}(k_{im})}\ 100 \qquad \frac{\sum_i (p_{im})\ k_{im}}{\sum_i (p_{im})\ k_{i0}}\ 100$$

Lehen egin genuen bezala, zenbakitzailearen batugai bakoitzean, oinarri-denborako prezio edo kopuruaz biderkatzen eta zatitzen bada, (**1**) formuletatik datozen (**2**) formuletara iritsiko gara.

$$\frac{P_t^p}{100} \qquad \frac{P_t^k}{100}$$

$$\frac{\sum_i \frac{p_{i1}}{p_{i0}}(p_{i0}\ k_{i1})}{\sum_i (p_{i0}\ k_{i1})}\ 100 \qquad \frac{\sum_i \frac{k_{i1}}{k_{i0}}(p_{i1}\ k_{i0})}{\sum_i (p_{i1}\ k_{i0})}\ 100$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\sum_i \frac{p_{it}}{p_{i0}}(p_{i0}\ k_{it})}{\sum_i (p_{i0}\ k_{it})}\ 100 \qquad \frac{\sum_i \frac{k_{it}}{k_{i0}}(p_{it}\ k_{i0})}{\sum_i (p_{it}\ k_{i0})}\ 100 \qquad \textbf{(2)}$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\sum_i \frac{p_{im}}{p_{i0}}(p_{i0}\ k_{im})}{\sum_i (p_{i0}\ k_{im})}\ 100 \qquad \frac{\sum_i \frac{k_{im}}{k_{i0}}(p_{im}\ k_{i0})}{\sum_i (p_{im}\ k_{i0})}\ 100$$

Ikusten denez, (**2**) formula hauek ere, indize sinpleen bidez konplexuak kalkulatzera eramaten gaituzte.

LASPEYRES eta PAASCHE-ren indizeak (**1**) formulen bidez batezbesteko gehikortu ponderatu bezala adierazten dira, eta (**2**) formulen bidez batezbesteko aritmetiko ponderatu bezala.

VI.3.1.3. FISHER-en indizeak

FISHER-en indize konplexu edo "indize ideal" deritzona, LASPEYRES eta PAASCHE-ren indizeen batezbesteko geometrikoa da.

Hots:

$$F_t^p = \sqrt{L_t^p \; P_t^p} \quad \text{FISHER} - \text{en prezioen indize konplexua.}$$

$$F_t^k = \sqrt{L_t^k \; P_t^k} \quad \text{FISHER} - \text{en kopuruen indize konplexua.}$$

VI.3.1.4. Propietate eta erlazio batzuk: bateragarritasuna eta alderantzizkotasuna.

Bateragarritasuna

Indizeen propietate garrantzitsu bat **bateragarri izatea da**, alegia, $B = P \cdot K$ berdintasuna betetzea B, P eta K balio, prezio eta kopuruen indizeak izanik.

LASPEYRES eta PAASCHE-ren indizeek ez dute betetzen erlazio hau, baina bai FISHER-enak.

Hots:

$$L_t^p \cdot L_t^k \neq 100 \; B_t$$

$$F_t^p \cdot F_t^k = 100 \; B_t$$

$$P_t^p \cdot P_t^k \neq 100 \; B_t$$

LASPEYRES eta PAASCHE-ren konbinaketaz ere betetzen da.

Hots: $L_t^p \cdot P_t^k = 100 B_t \qquad P_t^p \cdot L_t^k = 100 \; B_t$

Sinboloen ordez formula osoak ipintzen baditugu, aurreko erlazioak erraz frogatuko ditugu.

Alderantzizkotasuna

t denborako o denborarekiko indizea eta o denborako t denborarekiko indizea elkarren artean balio alderantzizkoak izateari deritzogu alderantzizkotasuna.

Indize sinpleek ondoko erlazio hau betetzen dute:

$$\frac{I_{t,0}}{100} \neq \frac{1}{I_{t,0} / 100}$$

LASPEYRES eta PAASCHE-ren prezio edota kopuru-indizeek ez dute betetzen, baina bai FISHER-enek.

$$\frac{L^{p}_{t,0}}{100} \neq \frac{1}{L^{p}_{t,0} / 100}$$

$$\frac{P^{p}_{t,0}}{100} \neq \frac{1}{P^{p}_{t,0} / 100}$$

$$\frac{F^{p}_{t,0}}{100} = \frac{1}{F^{p}_{t,0} / 100}$$

LASPEYRES eta PAASCHE-ren prezio edota kopuru-indizeen konbinaketaz ere betetzen da.

$$\frac{L^{p}_{t,0}}{100} = \frac{1}{P^{p}_{t,0} / 100} \qquad \frac{P^{k}_{t,0}}{100} \neq \frac{1}{L^{k}_{t,0} / 100}$$

VI.3.1.5. Kalkulua

Ondoko taulan labore batzuen prezio eta kopuruak ditugu, 1.960 eta 1.964 urteetakoak hain zuzen.

	Arroza		Artoa		Garia	
	p	k	p	k	p	k
1.960	24	100	18	40	20	50
1.964	32	80	12	40	40	70

LASPEYRES-en prezioen indizeaz:

$$\sum_i p_{it}k_{i0} = 32(100) + 12(40) + 40(50) = 5.680$$

$$\sum_i p_{it}k_{i0} = 24(100) + 18(40) + 20(50) = 4.120$$

$$L_{64}^{p} = \frac{5.680}{4.120}100 = 138$$

Alegia, oinarri-denborako kopuruaren balioa, prezioen bidez %38 gehitu dela, baina gehienetan, "prezioak 1.960-1.964 tartean, %38 inguruan goratu direla" esango da.

Bestalde, PAASCHE-ren prezio-indizeaz:

$$\sum_i p_{it}k_{it} = 32(80) + 12(40) + 40(70) = 5.840$$

$$\sum_i p_{io}k_{it} = 24(80) + 18(40) + 20(70) = 4.040$$

$$P_{64}^{p} = \frac{5.840}{4.040}100 = 144$$

Indize honek LASPEYRES-enak bezalako esanahia du, ordea, eraiketa desberdina duenez, oraingo emaitza %44 da.

VI.4. INDIZE KONPLEXUEN ERAIKETAN SORTZEN DEN ZENBAIT ERAGOZPEN

VI.4.1. Aldagaien hautapena

Dakigun bezala, aldagaien multzo bat batera aztertzea, indize konplexuari dagokio.

Beraz, aipatu dugun multzoaren hautapena, lehendabiziko arazoa izango da. Adibidez, aldagaiak nekazal prezioak baldin badira, beren kopurua nahiko handia izango da, areago, kalitate desberdinak kontuan hartzen baditugu; beraz, gehienetan, ondasun garrantzitsuenetatik azpipopulazio bat hautatuko da.

Honek, produktuen eta beren kalitateen izendatze zehatz batetara eramaten gaitu, honela, denbora iragatean, datu-segidak erizpide berdinekoak izango baitira.

VI.4.2. Somatutako leku eta denboraren hautapena

Somatutako zenbakizko datuak hartzean arduraz produktuen kalitateak definitzea eta somaketen leku eta denborak jakinaraztea komeni da. Prezioen kasuan, lekutzat produkzio, salketa edo kontsumo-lekua har liteke; denboratzat une bat (aste, egun edo hilabete bat) har liteke. Ohituraz, maiztasun, adiera edo garrantzi gehieneko erizpideari jarraituz aukeratzen dira leku eta denborak; beraz diren kasu guztien arteko hautapen aleatorio bat ez da posible.

VI.4.3. Talde eta azpitaldeen hautapena

Indize konplexua kalkulatzean, honetaz gainera aldagai, talde eta azpitalde batzuren “konplexua” kalkulatzen badugu, argibide gehiago izango dugu.

Adibidez, nekazal prezioen arloan bi talde handi kontsidera genitzake, sekain lurrena eta lur ureztatuena. Lehenengo taldean azpitaldeak har litezke, adibidez, laboreak, lekariak..., eta bigarrenean, adibidez, barazkiak, fruituak... Beste sailkapen batetan, barne-kontsumorako eta esportaziorako produktuak izan zitezkeen.Talde eta azpitalde hauen hautapena, egin nahi den ikerketaren menpe egongo da.

VI.4.4. Oinarri-denboraren hautapena

Oinarriari dagozkion datuak, beti ehun bezala hartzen dira, orduan, esan genuen bezala, garrantzitsuena “urte normal” bat hartzea da.

Indizearen barnean dauden aldagaiak, aldaketa nabarmenik ez badute (adibidez, industri produktuetan gertatzen den bezala), urte normal batetakoak oinarritzat har daitezke; bestalde, aldaketa handiak baldin badituzte (adibidez nekazal produktuetan bezala), hiruzpalau urtetako erdineurri bat hartuko da oinarritzat.

Beste arazo garrantzitsu bat zera da, oinarri-denbora oraingo denboratik ez dela oso urrun egon behar, baina, dena dela, beste galdera batetan (oinarri aldaketari buruzkoan) honetaz gehiago hitz egingo da.

VI.4.5. Formula eta ponderazioen hautapena

Formula eta ponderazioek elkarbide bat dute. Ponderazioen informazioa lortu ezin bada, normalki, indize sinplearen batezbesteko aritmetikoa egingo da, baina, lortu ahal bada, gehienetan LASPEYRES, PAASCHE eta FISHER-en indizeak erabiliko dira.

LASPEYRES-en indizean, ponderazioa oinarri-denborakoa denez, informazio gutxiena behar du; PAASCHE-renak gehiago behar du, ponderazioa aldakorra baita; eta azkenik, FISHER-ena, dakigunez, beste bien batezbesteko geometrikoa da.

VI.4.6. Indizearen esangura eta zabaldura

Ondasunen kopuruaren menpe zabaldura dago, adibidez, %75 edo %90 bat izango da. Batzutan, hautatu gabe gelditu diren ondasunetariko informazioa baldin badugu, zuzenketak sar daitezke indizeetan.

Indizearen esangura, batezbestekoen ikuspegitik begiratuko dugu; indize konplexuarekin batera sakabanatzearen neurri bat edukitzen dute eta hau handiago da oinarri-denboratik urruntzen garenean.

LASPEYRES eta PAASCHE-ren indizeak batera kalkulatzen baditugu, emaitzen desberdintasuna gehitu egingo da oinarri-denboratik urruntzen garenean eta hau sakabanatzearen aipamen bezala har dezakegu.

Desberdintasun hau nahiko handia egiten denean, oinarrien berriztapena egitea komeniko da.

VI.5. ERAGOZPEN BEREZI BATZU

VI.5.1. Oinarri-aldaketa indize sinpleetan

Adibide batez ikus dezagun.

Urtea	I_t(oinarria = 1.955)	I_t (oinarria = 1.965)
1.955	100	36'4
1.956	116'7	42'4
1.957	150	54'5
1.958	150	54'5
1.959	158'3	57'6
1.960	125	45'5
1.961	100	36'4
1.962	133'3	48'5
1.963	166'7	60'5
1.964	200	72'7
1.965	275	100

Salmenten indize bat baldin bada eta oinarria 1.965. urtean berritzen bada, momentu horretan denborazko segida hautsi egiten da eta denborazko segida berrian kalkuluak atzerantz egiten ditugu erregela proportzional baten bidez (I_t-ren zenbaki bakoitza bider 100 / 275 biderkatzean).

Alegia, 1.955 urteko salmentak %100 baldin baziren, orain 1.965 urtekoak %100 dira.

Bi segida hauek adierazten digute: %33'3 gehiago saldu zela 1.962. urtean 1.955.ean baino eta %51'5 gutxiago saldu zela 1.962. urtean 1.965.ean baino.

VI.5.2. Berriztapen eta loturak indize konplexuetan

Produkzio, aktibitate eta gastuen aldaketaren ondorio bezala, berriztapena inposatzen da. Momentu batetan ondasunik erabilgarrienak zirenak, bigarren postu batetan gelditzen dira denbora iragatean edo desagertu egiten dira. Bestaldetik ondasun berriak agertzen dira, gero arrunt bihurtzen direlarik.

Horregatik ondasun multzoa ezin dezakegu utz aldatu gabe. Denbora iragatean, aldagaiak, oinarriak eta ponderazioak aukeratu beharko ditugu berriro.

Indize konplexuen berriztapenak, segidaren hausturaren arazoa sortzen du.

Ikus dezagun nola egiten den lotura kasu erreal batetan:

Estatistika-Institutuak (Estatu espainolekoak), kontsumo-prezioen indizea berriztatu egiten zuen eta bi segida daude, bata 1.936. urtean oinarritua eta bestea 1.958.ean.

Kontsumo-prezioen indizea:

Urtea	Oinarria 1.936 = 100	Oinarria 1.958 = 100
1.956	643'1	—-
1.957	712'4	—-
1.958	807'7	100
1.959	866'7	—-
1.960	876'9	—-
1.961	—-	111'3
1.962	—-	117'3

Lotura-eragiketak segida bakar bat lortzen duenez gero, 1.936. urtean edo 1.958.ean oinarrituz egin liteke. Bietan indize sinpleetan ikusi genuen erregela proportzionalen bidez egingo da.

Bi segidak, bi indize desberdinak ditugun urtean (1.958 urtea kasu honetan) berdinak egiten dira. Oinarri bakartzat 1.936. urtea hartzen badugu, orduan 1.958.eko 100, bider 8'077 biderkatzean 807'7 bihurtzen da, eta honela egingo da segida osoaz; aldiz, oinarri bakartzat 1.958 urtea hartzen badugu, orduan, 1.936.ean oinarriturik dagoen segida, 8'077 balioaz zatituko da.

Ondoko taulan, loturik dauden bi segida ikusiko ditugu.

Urtea	Oinarria 1.936 = 100	Oinarria 1.958 = 100
1.956	643'1	79'5
1.957	712'4	82'2
1.958	807'7	100
1.959	866'7	107'3
1.960	876'9	108'6
1.961	899'0	111'6
1.962	947'2	111'4

Argiro, indize sinpleen alderantziz, bi segida hauek ez dira konparagarriak, aldagai eta ponderazio desberdinez eginak baitaude, baina, ez dugu beste biderik, indize konplexuari jarraitasuna eman nahi badiogu.

Ez dugu ahaztu behar, indizeak fenomeno baten eboluzioa ikusteko erabiltzen diren adierazle batzu besterik ez direla.

VI.6. ZENBAKI-INDIZEEN APLIKAPENAK

Indizeen formulak era orokor batean ikusi ondoren, bi multzotan sailka ditzakegu, prezio eta kopuruenak, baina, formula hauek, indize batzutan zehaztu egiten dira, adibidez, industri produkzioen indizeak, alokairuen indizeak, kontsumo-prezioen indizeak, balore mugigarrien indizeak, inportazioen indizeak, esportazioen indizeak eta abar, gehienak ekonomi aktibitatearen eboluzioa ikusteko baliagarriak direlarik.

Ikasgai honetan, kontsumo-prezioen indizea ikusiko dugu, garrantzitsuena baita.[1]

VI.6.1. Kontsumo-prezioen indizeak

Hipotesi bezala familia baten bizimaila konstantetzat hartzen badugu, denbora-une desberdinetan familia horrek kontsumoan ordaintzen duenaren gain konparaketak egitea da indize honen helburua. Alegia, aurrekontu (familia baten bizimaila deritzona) baten aldaketak neurtu nahi ditu.

Egikera erabiliena "erosketa-otarrea" deitzen dena da; hau inkesta baten bidez, Famili Aurrekontuen Inkesta, alegia, aurrekontuan sartzen diren ondasunen kopuruek, zerbitzuek eta industri produktuek osatzen dute.

Hartzen den familia, "normala" da, hots, bere aurrekontua batezbestekoari hurbiltzen zaio.

Oinarri-denborako prezioaz eta oraingo prezioaz "otarre"-aren balioa kalkulatzea, izango da arazo bakarra, eta bi hauen zatidura kontsumo-prezioen indizea da.

Izan bitez:

$k_{10}\ k_{20}\ \ldots\ k_{n0}$ "otarre"-ko ondasunen kopuruak (bizimaila presuposatzen dutenak).

$p_{10}\ p_{20}\ \ldots\ p_{n0}$ haien prezioak oinarri-denboran.

eta

$p_{11}\ p_{21}\ \ldots\ p_{n1}$ haien prezioak oraingo unean.

---

1. Besteak UEU-k 1979. urtean argitaratu zuen ZENBAKI-INDIZEAK liburuxkan ikus daitezke.

Orduan **kontsumo-prezioen indizea,** ondoko formulaz edukiko dugu.

$$I_{1,0} = \frac{\sum_{i=1}^{n} p_{i1}k_{i0}}{\sum_{i=1}^{n} p_{i0}k_{i0}} 100$$

Ikusten denez, indize hau LASPEYRES-en indize bat da.

Lehen esan dugunez, denbora iragatean indizearen oinarria aldatuko ez balitz, eta honekin batera indizearen estruktura gaurkotuko ez balitz, indizea balio gabe geratuko litzateke. Zer esanik ez indizea kontsumo-prezioena izanik, azken urteetan jasaten ditugun aldaketa teknologiko eta ekonomikoen egoeraz baldintzatuta dagoela.

Espainiako Estatistika Institutuak KPI kalkulatzean oinarri hauek finkatu ditu: 1936, 1958, 1976, 1983 eta 1985.eko uztaila.

Azken urte hauetako egoera, hauxe da: Famili Aurrekontuen Inkesta berria 1990/91 urteetan burutu zen, oinarri aldaketa 1992 urtea oinarritzak harturik egin zelarik. Kontsumo Prezioen Indize berria 1993. urteko urtarriletik aurrera ateratzen da.

Dakusagunez lehen aldaketa 22 urte iragan ondoren izan zen, eta bigarrena 18 urte iragan ondoren; azkeneko urteetan gero eta hurbilago dago aldaketa, kontsumoaren estrukturaren aldaketak beharturik.

VI.6.2. Moneta-unitate arruntetan dauden magnitudeen deflazioa

Zenbaki-indizeen aplikazio garrantzitsuenetako bat deflazioa da.

Prezio edo balioen moneta-segida bat deflaktatzea, segida hori moneta efektuaz askatzea da.

$\sum_i p_{it}k_{it}$ balio-segida, moneta-balio arruntetan dago eta balio errealetan edo moneta-balio konstanteetan ipini nahi badugu, deflaktatze eragiketa eginez $\sum_i p_{i0}k_{it}$ balio-segidan bihurtuko da.

Bidea, PAASCHE-ren prezio-indizeaz zatitzea izango litzateke.

Honela

$$\frac{\sum_i p_{it}k_{it}}{P_t^p} = \frac{\sum_i p_{it}k_{it}}{\frac{\sum_i p_{it}k_{it}}{\sum_i p_{i0}k_{it}}} = \sum_i p_{i0}k_{it}$$

P_t^p ez daukagunean, beste "ad hoc" indizea hartuko da, baina segida desberdin batzu parekatu nahi baditugu, indize berbera erabiliko dugu.

Adibidea: Ondoko taulan 90-92 urte-tarterako ondasun baten salmentak pezeta arruntetan ditugu. Zein izango litzateke salmenta horien balioa pezeta konstanteetan, urte horietan Laspeyres-en prezio-indizeak: 100, 131.42, 162.43, hain zuzen, deflaktatzeko erabiltzen baditugu?

Urtea I_t(oinarria = 1.955) I_t (oinarria = 1.965)

Urtea	Salmentak pta arrut.	Deflazio--indizea	Salmentak 1990.eko pta. konstanteetan
1990	1500000	100	1500000
1991	1800000	131.42	1396654
1992	2100000	162.43	1300873

VII. DESKRIBAPEN ESTATISTIKOAREN ADIERAZPIDE GEOMETRIKOAK

VII.1. OROKORTASUNAK

*VII.2. **X** DATU-MATRIZEAREN BI ADIERAZPIDE GEOMETRIKO*

VII.3. BI ALDAGAI ETA BI OHARPEN

VII.4. BI ALDAGAI ETA HIRU OHARPEN

VII.5. ERREGRESIO BAKUNA ALDAGAI ESTATISTIKOEN BALIO-MULTZOEN ESPAZIOAN

VII.1. OROKORTASUNAK

Ikasgai honetan batezbestekoa, bariantza eta kobariantza geometrikoki ikustea hartzen dugu helburutzat.

Matrize-estatistikoak azaltzen genuen ikasgaian, estatistika anizkoitzean ikasketak matrizialki egitea beharrezko bihurtzen dela esaten genuen.

Errealitatea estatistikoki[1] ikasi nahi badugu, bada, formulazio matematiko bektoriala edota matriziala nahitaezkoak eta bide batez paregabeak direla esan daiteke, errealitatea bektoriala baita.

$\mathbf{X}_{(m\cdot n)}$ datu-matrizearen bektoreak, $\mathbb{R}^n$, $\mathbb{R}^m$ espazioetako bektoreak alegia, aurkezten hasiko gara; ondoren, $\mathbb{R}^2$, $\mathbb{R}^3$ espazioetan zentraturik, aipatutako estatistikoak geometrikoki ikusiko ditugu.

Zoritxarrez n edo m 3 baino handiago direnean iruditan ezer gutxi ikusiko dugu, baina, $\mathbb{R}^2$, $\mathbb{R}^3$ espazioak ondo aztertu baditugu espazio anizkoitzean gertatzen dena errazago ulertuko dugu.

VII.2. X DATU-MATRIZEAREN BI ADIERAZPIDE GEOMETRIKO

Lehen aurkeztutako $\mathbf{X}_{(m\cdot n)}$ datu-matrizean bi bektore mota kontsidera daitezke:

1) $\mathbb{R}^n$ espazioan, indibiduoek dituzten n koordenatuek n aldagaien arabera $\mathbb{R}^n$ espazioko bektoreak osatzen dituzte.

2) Bestalde, $\mathbb{R}^m$ espazioan, aldagaiek dituzten m koordenatuek m indibiduoen gain $\mathbb{R}^m$ espazioko bektoreak osatzen dituzte.

---

1. Estatistika Deskribatzailean kokaturik, eta konkretuki, Datu-Analisiaren atarian, hau da, probabilitate-indukzioan oinarritzen den Inferentzia Estatistikotik at.

Hots:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \cdots & x_{1j} \cdots & x_{1n} \\ x_{21} & x_{22} \cdots & x_{2j} \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} \cdots & x_{ij} \cdots & x_{in} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} \cdots & x_{mj} \cdots & x_{mn} \end{bmatrix}$$

Errenkadak, $\mathbb{R}^n$ espazioko bektoreak dira, i. indibiduorako edo oharpenerako (x_{i1} x_{i2} ... x_{ij} x_{in}) koordenatuak direlarik. Errenkadetan, bada, m **indibiduo-puntu** edo oharpen-puntu ditugu, orain arte edozein hodeitan genituen puntuak alegia.

Zutabeak, aldiz, $\mathbb{R}^m$ espazioko bektoreak dira j. aldagairako (x_{1j} x_{2j} ... x_{ij} ... x_{mj}) koordenatuak direlarik. Zutabetan n **aldagai-puntu** ditugu hemendik aurrera kontutan hartuko dugun puntu-hodei berri bat osatzen dutelarik.

VII.3. BI ALDAGAI ETA BI OHARPEN

Bi aldagai bi indibiduoen gain neurtu ondoren honako emaitza hauek ditugu.

$$\begin{matrix} & X_1 & X_2 \\ \omega_1 \\ \omega_2 \end{matrix} \begin{bmatrix} x_1(1) & x_2(1) \\ x_1(2) & x_2(2) \end{bmatrix} = \text{(konkretuki)} = \begin{bmatrix} 1 & 5 \\ 3 & 2 \end{bmatrix} \text{ datu-matrizea}$$

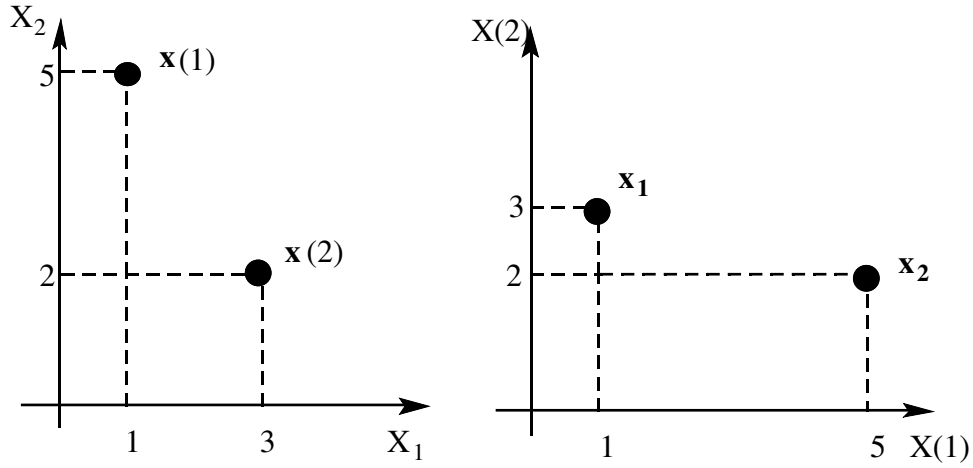
adibide bezala erabiliko dugu.

Kasu honetan $\mathbf{x}(1) = (1,5)^T$, $\mathbf{x}(2) = (3,2)^T$, $\mathbb{R}^2$-ko bi bektoreak indibiduo--puntuak edo oharpen-puntuak dira. $\mathbf{x}_1 = (1,3)$, $\mathbf{x}_2 = (5,2)$, $\mathbb{R}^2$–ko bi bektoreak, aldiz, aldagai-puntuak.

Konturatu behar dugu datu-matrizeko bektoreak ez direla desberdintzen errenkada-bektoreak edo zutabe-bektoreak izateagatik bakarrik, daukaten desberdintasuna benetan garrantzitsuagoa da: lehenengoak balio-multzo ez--homogenoak dira indibiduo baten gain aldagai desberdinen neurriak izanik; bigarrenak, ordea, balio-multzo homogenoak dira aldagai bakoitzaren neurriak izatean.

Datu-matrizearen (i,j) elementua adierazteko bi ikur ezberdin aukeratu ditugu. Hauexek dira: $x_{ij} = x_j(i)$

Biek X_j aldagaiak ω_i indibiduoaren gain hartzen duen balioa adierazten digute. Dakusagun, bada, adierazpide grafikoak:



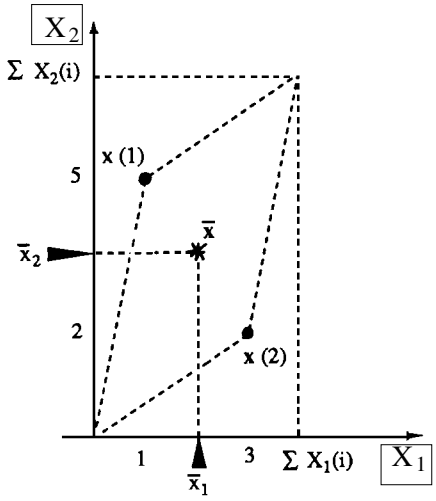
Indibiduoen Hodeia Aldagaien Hodeia

X_1, X_2, aldagaien balioen batezbestekoek betebehar desberdinak betetzen dituzte bi hodei horietan

$$\bar{x}_1 = \frac{1}{m}\sum_i x_1(i) = 2 \qquad \bar{x}_2 = \frac{1}{m}\sum_i x_2(i) = 3'5$$

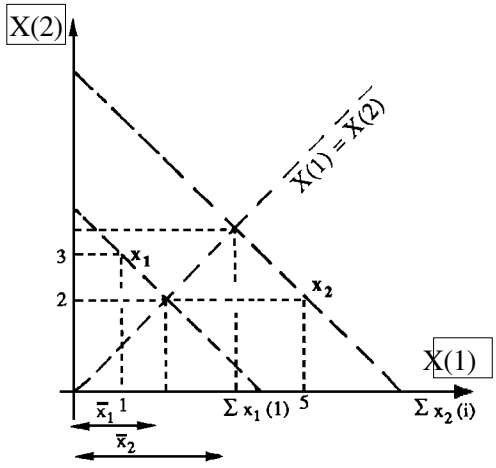
INDIBIDUOEN HODEIAN

Indibiduo-puntuen grabitate-zentrua finkatzen dute.



ALDAGAIEN HODEIAN

Aldagai-puntuen osagaien batezbestekoa



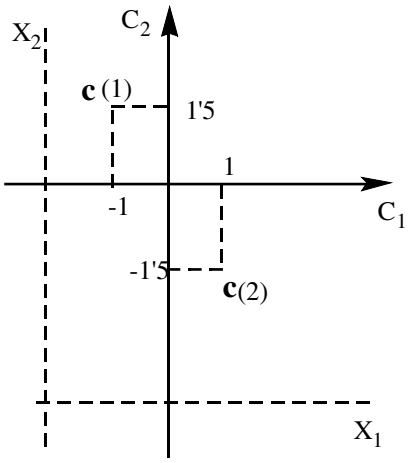
Aldagai-bektorearen batezbestekoa, X(1) = X(2) erdikarian aldagai-puntuak daukan proiektatutako puntuaren osagai bakoitza bezala ikusten da.

Aldagaien balio bakoitzaren **zentratze-eragiketa**:

C_i (i) = X_j (i) - $\bar{x}_j$ zentratze-eragiketa eginez, **C** datu-matrize zentratua lortzen dugu.

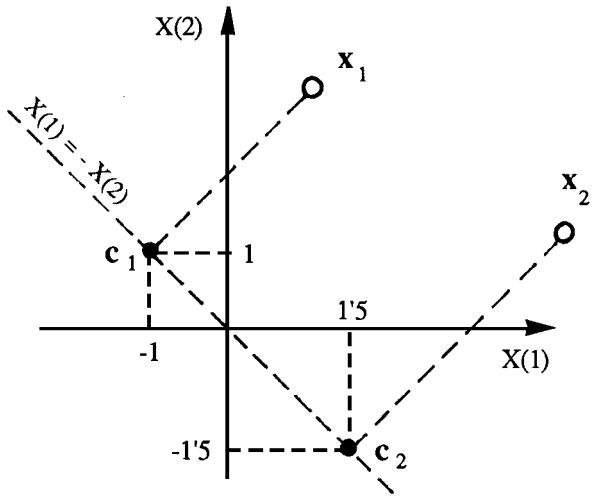
$$C = \begin{bmatrix} c_1(1) & c_2(1) \\ c_1(2) & c_2(2) \end{bmatrix} = \begin{bmatrix} -1 & 1,5 \\ 1 & -1,5 \end{bmatrix}$$

INDIBIDUOEN HODEIAN



Zentratzeak ardatzen aldaketa paralelo bat egitea suposatzen du hodei honetan.

ALDAGAIEN HODEIAN



Hodei honetan, ordea, batezbestekoz eta baturaz zero den $X(1) = -X(2)$ erdikarian aldagai-puntuak proiektatzea da zentratze-eragiketak suposatzen duena.

Aldagai batek bi oharpen dituenean, zentratzea $X(1) = -X(2)$ erdikarian proiektatzea da, bi balio zentratuak, berdinak balio absolutuz eta aurkakoak baitira. Ikus grafikoan.

Horrexegatik X(2 x 2) datu-matrizeak ezaugarri oso bereziak ditu.

Dakusagun:

1) Desbidazio standarda aldagai zentratuaren koordenatuen balio absolutua da.

$$S_j = |c_j(1)| = |c_j(2)|$$

Kasu honetan: $S_1 = |-1| = |1| = 1$

$S_2 = |1'5| = |-1'5| = 1'5$

Dakigunez:

$$S_1^2 = \frac{1}{m}\sum_i c_1^2(i) \qquad S_2^2 = \frac{1}{m}\sum_i c_2^2(i)$$

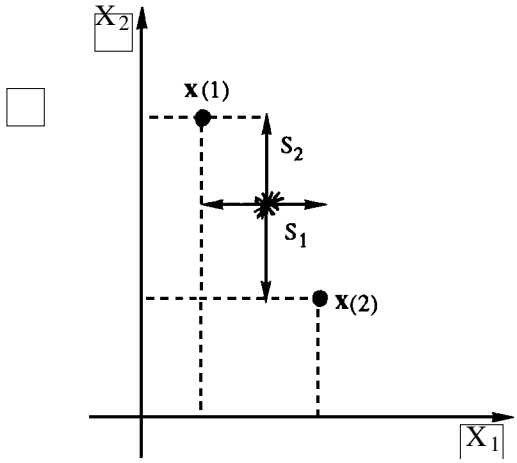
$$S_1 = \frac{\|c_1\|}{\sqrt{m}} \qquad S_2 = \frac{\|c_2\|}{\sqrt{m}}$$

m = 2, hots bi indibiduo edo bi oharpen izanik

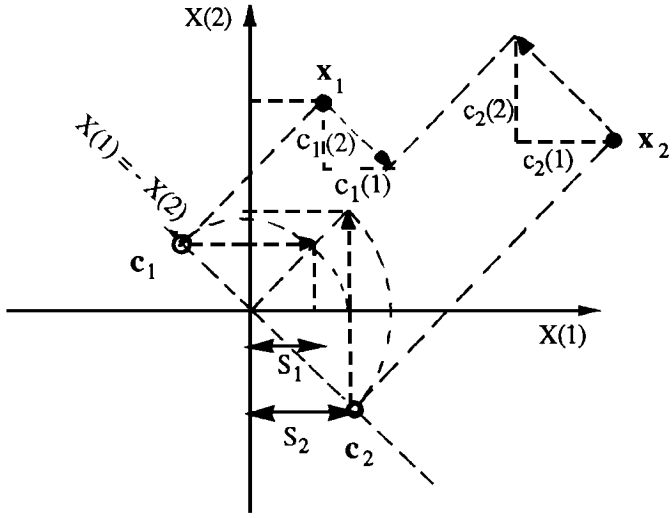
$$\mathbf{L} = \begin{bmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ 1'5 & -1'5 \end{bmatrix}\begin{bmatrix} -1 & 1'5 \\ 1 & -1'5 \end{bmatrix} = \begin{bmatrix} 1^2 & -1.1'5 \\ -1.1'5 & 1'5^2 \end{bmatrix}$$

Kobariantza matrizea lortzean, matrize hauen bigarren berezitasuna agertzen zaigu, baina horretaz geroxeago hitz egingo dugu.

Dakusagun, orain bada, desbidazio standarda nola agertzen den indibiduoen hodeian.



Kasu berezi honetan (**X**(2 x 2)) desbidazio standarda edo tipikoa "desbidazioa" balio absolutuan da.



X(1) = X(2) erdikarian aldagai-puntua proiektatzean, proiektatutako puntuaren koordenatu bakoitza batezbestekoa zen bezalaxe, orain erdikarian aldagai-puntu zentratua proiektatzean, proiektatutako puntuaren koordenatu bakoitza desbidazio standarda da.

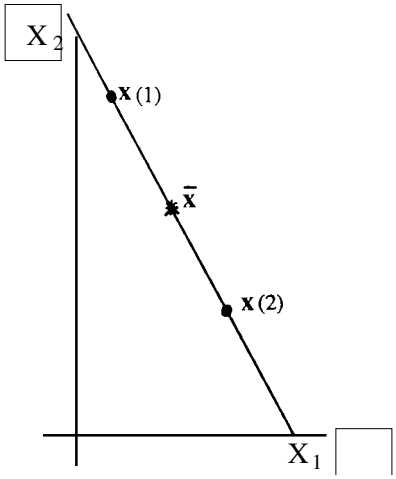
$\mathbf{C}_j$ bektoreak eta bere proiekzioak norma bera daukate:

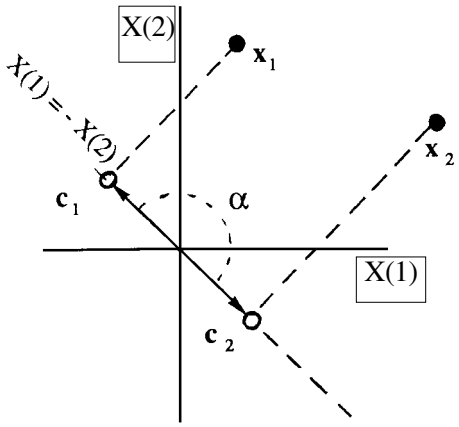
$$\left\| c_j \right\| = \sqrt{m}\ S_j$$

Aldagai zentratuaren norma sakabanatzearen neurri on bat da, askotan erabilia izanik.

Azkenik azpimarratu behar dugu aldagai zentratuaren norma, jatorrizko aldagaiak erdikarira daukan distantzia euklidearra dela.

2) Bi oharpen ditugunean aldagaiak guztiz korrelatuak dira beti, hots koerlazio-koefizientea 1 edo -1 (gure adibidean bezala) izango da. Ikus **L** aurreko orrialdean.





$$r_{12} = \frac{S_{12}}{S_1 S_2} = \frac{< c_1, c_2 > / m}{\frac{\|c_1\|}{\sqrt{m}} \frac{\|c_2\|}{\sqrt{m}}} = \frac{< c_1, c_2 >}{\|c_1\| \, \|c_2\|} = \cos \alpha$$

Dakusagunez, aldagai zentratuen azpiespazio bektoriala X(1) = -X(2) zuzena da, horrexegatik edozein aldagai bikotek osatzen duen angelua zero edo ehun eta laurogei gradukoa izango da.

Horrek, dakigunez, erregresio perfektua halabehartzen du.

Indibiduo-puntuek ezin dutela zuzen baten gainean ez izan ikusten dugu ezkerraldeko hodeian, lerrokatuak daude.

Aldagai zentratu bakoitza bestearen eskalarra dela ikusten dugu, ordea, eskuinaldekoan.

Ondoren, hau baino errepresentagarriago den kasu bat aztertuko dugu, bi aldagai eta hiru oharpenen kasua, alegia.

Ariketa: **X** datu-matrizea $\begin{bmatrix} 1 & 4 \\ 5 & 6 \end{bmatrix}$ matrizea izanik. Ikus ezazu, grafikoki, $\mathbf{c}_1$, $\mathbf{c}_2$ bektoreek zero graduko angelua osatzen dutela edota $r_{12} = 1$ dela.

VII.4. BI ALDAGAI ETA HIRU OHARPEN

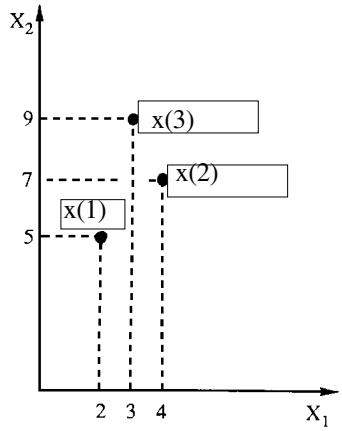
Ondoan, bi aldagai eta hiru indibiduoren kasuko matrizea daukagu.

$$\mathbf{X} = \begin{bmatrix} x_1(1) & x_2(1) \\ x_1(2) & x_2(2) \\ x_1(3) & x_2(3) \end{bmatrix}$$

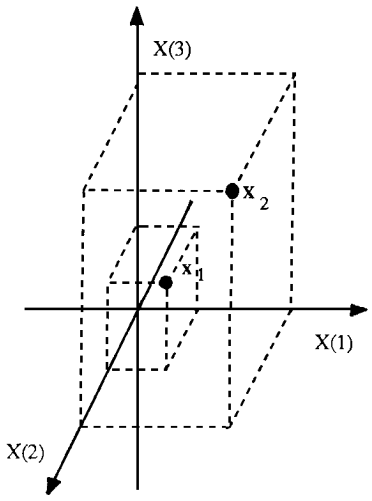
Eta adibide bezala kontsideratzen duguna.

$$\mathbf{X} = \begin{bmatrix} 2 & 5 \\ 4 & 7 \\ 3 & 9 \end{bmatrix}$$

Indibiduo-puntuen hodeia



Aldagai-puntuen hodeia

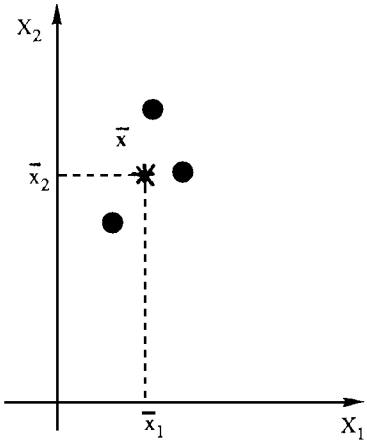


$$\bar{x}_1 = \sum_i x_1(i) / m = 3$$

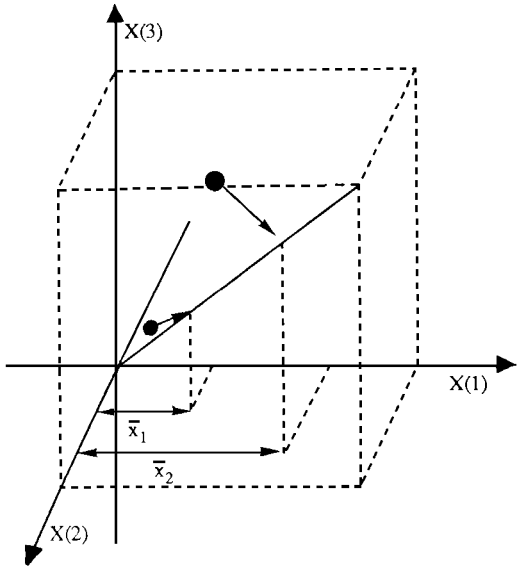
$$\bar{x}_2 = \sum_i x_2(i) / m = 7$$

Hots $\bar{\mathbf{x}} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}$

Dakusagun, bada, adierazpide grafikoak:



Batezbesteko bektorea indibiduoen barizentrua edo grabitate-zentrua da.



Aldagai bektorearen batezbesteko osagaia X(1) = X(2) = X(3) erdikarian proiektatutako puntuaren osagai bakoitza da.

Zentratze-eragiketa:

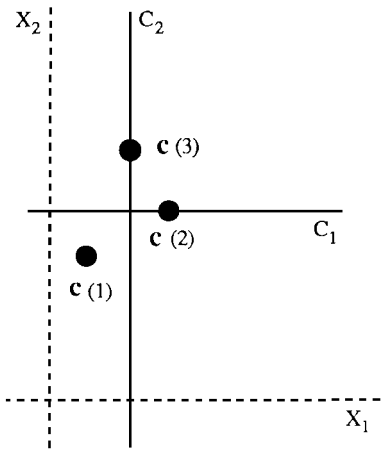
$$\bar{x}_1 = \sum_i x_1(i) / m = 3$$

$$\bar{x}_2 = \sum_i x_2(i) / m = 7$$

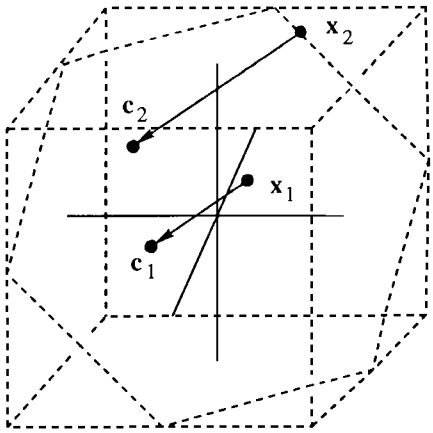
Hots $\bar{\mathbf{x}} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}$

datu-matrize zentratua lortzen dugu.

Grafikoki:



Berriro ere ardatzen aldaketa paralelo bat daukagu.

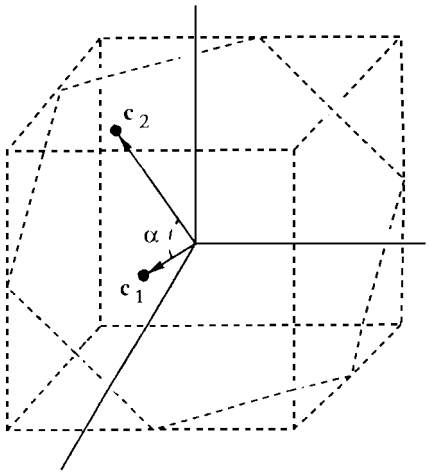


Lehen X(1) = -X(2) erdikarian proiektatzen ziren aldagai-puntuak, hots, baturaz eta batezbestekoz zero diren bektoreen azpiespazioan.

Orain, bada, berdin gertatzen da baina kasu honetan baturaz eta batezbestekoz zero diren bektoreen azpiespazioa plano bat da, X(1) + +X(2) + X(3) = 0 planoa alegia.

Irudian, kubo baten barruan dagoen hexagonoaren bidez ikusarazi nahi da azpiespazio hori, hau da, hexagonoak X(1) + X(2) + X(3) = 0 planoa ikusarazteko besterik balio ez duelarik.

Aldagai zentratuen grafika handituz aldagaien arteko koerlazioa geometrikoki ikusiko dugu.



Ikusi genuenez:

$$\| c_1 \| = \sqrt{m} \quad S_1$$

$$\| c_2 \| = \sqrt{m} \quad S_2$$

$$r_{12} = \frac{< c_1, c_2 >}{\| c_1 \| \| c_2 \|} = \cos \alpha \qquad \alpha = \text{arc} \cos r_{12}$$

Aldagaien osagaiak ez dira berdinak balio absolutuan ez eta aldagaiek ez dute zergatik guztiz korrelatuak izan behar ere.

Darabilgun adibidean **L** eta **R** matrizeak hauek dira:

$$\mathbf{L} = \begin{bmatrix} 2/3 & 2/3 \\ 2/3 & 8/3 \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

hots $\alpha = 60^{\circ}$

$\alpha = \text{arc cos } 1/2$

Ariketa: $\mathbf{c}_2$ bektorearen ordez direkzio berean dagoen $\mathbf{c}_2$ = (-1, 0, 1) bektorea kontsideratuz, ikus ezazu $\mathbf{c}_1$, $\mathbf{c}_2$ bektoreen arteko angelua 60ºkoa dela.

Oharra: $\mathbf{c}_1$, $\mathbf{c}_2$, 0 puntuek alde berdineko triangelu bat osatzen dute, alde bakoitzak 2 balio duelarik.

DATU-MATRIZE **TIPIFIKATUA** ETA **ZENTRU-NORMALIZATUA**

Dakigunez aldagaien balioak tipifikatuak edo standardizatuak dituen datu-matrizeari tipifikatua deritzogu.

Adibidean:

$$\mathbf{C} = \begin{bmatrix} -1 & -2 \\ 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad S_1 = \sqrt{\frac{2}{3}}, \quad S_2 = 2\sqrt{\frac{2}{3}} \quad \text{izanik}$$

$$\mathbf{T} = \left[\frac{x_j(i) - \bar{x}_j}{S_j} \right] = \begin{bmatrix} \frac{\sqrt{3}}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{2}} \\ \frac{\sqrt{3}}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix}$$

Era berean, aldagai zentratuen balioak dagozkien normaz zatituz, datu-matrize zentru-normalizatua lortuko dugu.

Hau da:

$$\|\mathbf{c}_1\| = S_1 \sqrt{3} = \sqrt{2}$$

$$\|\mathbf{c}_2\| = S_2 \sqrt{3} = 2\sqrt{2}$$

$$\mathbf{N} = \left[\frac{x_j(i) - \bar{x}_j}{S_j \sqrt{m}} \right] = \begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Aldagai-puntu tipifikatuen norma $\sqrt{m}$ da.

Hots:

$$\left\| t_j \right\| = \left\| \frac{x_j(i) - \bar{x}_j}{S_j} \right\| = \left\| \frac{\mathbf{c}_j}{S_j} \right\| = \frac{\left\| \mathbf{c}_j \right\|}{S_j} = \frac{\left\| \mathbf{c}_j \right\|}{\left\| \mathbf{c}_j \right\| / \sqrt{m}} = \sqrt{m}$$

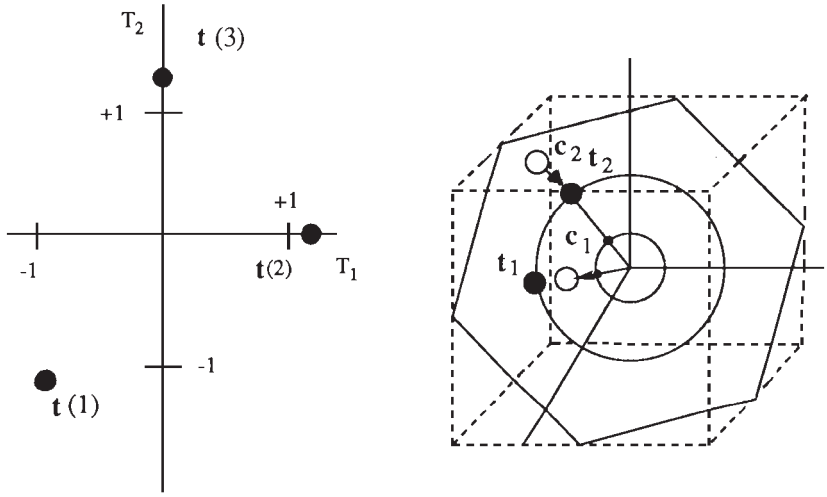
Noski, $\mathbf{N}$ matrizeko zutabe-bektoreek 1 daukate normaz, bektoreak normalizatuak baitaude.

Edozein kasutan, aldagaiak zentratzeak translazio bat bakarrik suposatzen badu ere, tipifikatzeak edota zentratu ondoren normaltzeak eskala aldatzea suposatzen du.

Horrexegatik, $\mathbb{R}^3$ espazioan aldagai-puntu tipifikatuak erradioz $\sqrt{3}$ duen zirkulu batetan eta $m \geq 4$ denean erradioz $\sqrt{m}$ duen esfera edo hiperesfera batetan kokatzen dira hurrenez hurren.

Aldagai-puntu zentru-normatuak, berriz, erradioz 1 neurtzen duen zirkulu, esfera, edo hiperesferan kokatzen dira.

Bi kasu horietan, balioak erlatibizatzen ditugu desbidazio standardaren arabera eta aldagai guztiak eskala berean ditugunez beraien direkzioak interesatzen zaizkigu.



Indibiduoen balioak tipifikaturik ditugu ezkerraldeko hodeian.

Eskumaldeko hodeian erradioz $\sqrt{3}$ duen zirkunferentzian aldagai-bektore tipifikatuak ikusten ditugu, baita, aldagai-bektore normatuak erradioz 1 duen zirkunferentzian ere.

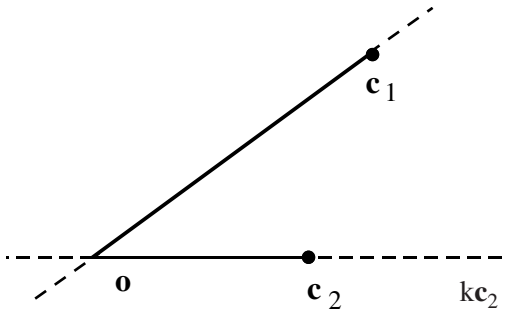
VII.5. ERREGRESIO BAKUNA ALDAGAI ESTATISTIKOEN BALIO--MULTZOEN ESPAZIOAN

X.2. atalean ikusi dugun bezala **X**(m x n) datu-matrizearen zutabeak $\mathbb{R}^m$ espazioko aldagai-bektoreak dira, m indibiduoen gain aldagai bakoitzaren balioak hartu baititugu.

Aldagai-bektore zentratuak, ordea, beraien osagaien edo balioen batura zero denez, $\mathbb{R}^{m-1}$ espazioko bektoreak dira. Konkretuki, X.3. eta X.4. ataletan, $\mathbb{R}^1$ eta $\mathbb{R}^2$ espazioetan kokaturik ikusi ditugun bi indibiduoen eta hiru indibiduoen kasuetarako, hurrenez hurren.

Bi aldagaien arteko erregresio bakuna (aldagai independente edo erregresore bat) beti, plano batean adieraz daiteke, alegia, $\mathbb{R}^{m-1}$ dimentsioko bi aldagai--bektore zentratuek sortarazten duten espazio bidimentsionalean.

Bi aldagai eta m oharpen suposaturik, ikus ditzagun, bada, X(m x 2) datu--matrizeko bi aldagai zentratuak beraiek osatzen duten planoan adieraziak.



X_1 aldagaiaren X_2 aldagaiarekiko erregresioa suposatuz eta aldagai--bektore zentratuen bidez adieraziz:

$$\hat{c}_1 = b_{12} c_2$$

Hau da, aldagaiaren balio estimatuen bektorea $\mathbf{c}_2$ bektorearen eskalarra da.

b_{12} $\mathbf{c}_2$, bada, O $\mathbf{c}_2$ segmentua kokatuta dagoen zuzenean kokatuko da, zuzen hau $\mathbf{c}_2$ bektorearen eskalar guztien leku geometrikoa edo k$\mathbf{c}_2$ direkzioa baita.

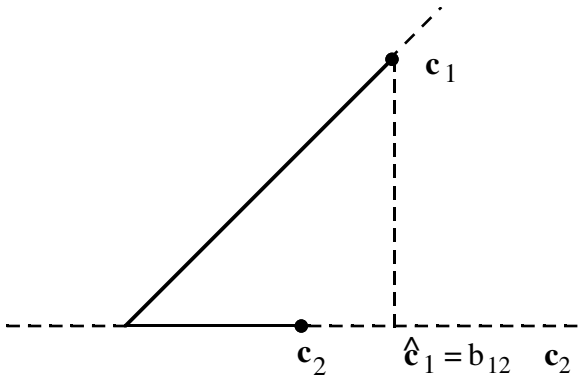
$\mathbb{R}^m$ espazioan, $\mathbf{c}_1$ eta b_{12} $\mathbf{c}_2$ bektoreen arteko distantzia, norma euklidearraz kalkula daiteke.

$$d = + \sqrt{\sum_i \left[c_1(i) - b_{12}\, c_2(i) \right]^2} = \sqrt{\sum_i \left[c_1(i) - \hat{c}_1(i) \right]^2}$$

Eta c_1 (i) - $\hat{c}_1$ (i) i. indibiduoaren hondarra edo errorea izanik, distantzia, errore karratuen baturaren erro karratua da, hau da, $\| \mathbf{e} \|$ norma. Balio hau txikiena, eta bide batez, S_e^2 hondar-bariantza txikiena izatea da erregresioaren helburua.

Erregresioa, bada, geometrikoki honela plantea daiteke:

$\hat{\mathbf{c}}_1 = b_{12}\, \mathbf{c}_2$ eta $\mathbf{c}_1$ balio-multzoen arteko distantzia txikiena izatea da bilatzen dena. Balio hau, dakigunez, $\mathbf{c}_1$ eta $\mathbf{c}_1$-ek, k$\mathbf{c}_2$ direkzioan, daukan proiekzio ortogonalaren arteko distantziari dagokio.



Honela, geometrikoki, b_{12} erregresio-koefizientea lor dezakegu:

proj. $\mathbf{c}_1 = b_{12} \| \mathbf{c}_2 \|$

eta dakigunez,

$$\text{proj. } \mathbf{c}_1 = \left\langle \mathbf{c}_1 \,\middle|\, \frac{\mathbf{c}_2}{\|\mathbf{c}_2\|} \right\rangle = \frac{\langle \mathbf{c}_1 \mid \mathbf{c}_2 \rangle}{\|\mathbf{c}_2\|}$$

orduan,

$$b_{12} = \frac{\langle \mathbf{c}_1 \mid \mathbf{c}_2 \rangle}{\|\mathbf{c}_2\|^2} = \frac{\langle \mathbf{c}_1 \mid \mathbf{c}_2 \rangle / m}{\|\mathbf{c}_2\|^2 / m} = \frac{S_{12}}{S_2^2}$$

Azkenik, balio estimatuen bektorea

$$\hat{\mathbf{c}}_1 = \frac{S_{12}}{S_2^2}\, \mathbf{c}_2$$

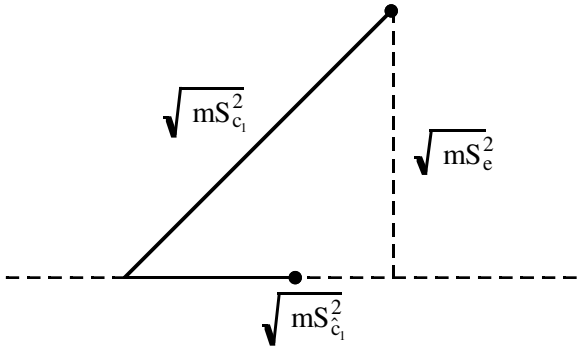
$\mathbf{c}_1$ bektorearen bektore-posizioaren osagaiak, $\hat{\mathbf{c}}_1$ bektorearekiko $\mathbf{c}_1$-ren $\mathbf{c}_2$-rekiko erregresioaren hondarrak dira.

Dakusagunez, $\mathbf{e}$ bektorearen direkzioa $\mathbf{c}_2$ eta $\hat{\mathbf{c}}_1$ bektoreen direkzioarekiko ortogonala da. Honela geometrikoki erregresioaren bi propietate hauek ikusten ditugu:

$-r_{ec_2} = r_{ex_2} = 0$, hau da, hondarrak eta aldagai independentea koerlazio gabeak dira.

$-r_{e\hat{c}_1} = r_{e\hat{x}_1} = 0$, hau da, hondarrak eta balio estimatuak koerlazio gabeak dira.

Azkenik, hain garrantzitsua den bariantzaren deskonposaketa batukorra, geometrikoki ikus dezakegu:



Pitagoras-en teorema aplikatuz,

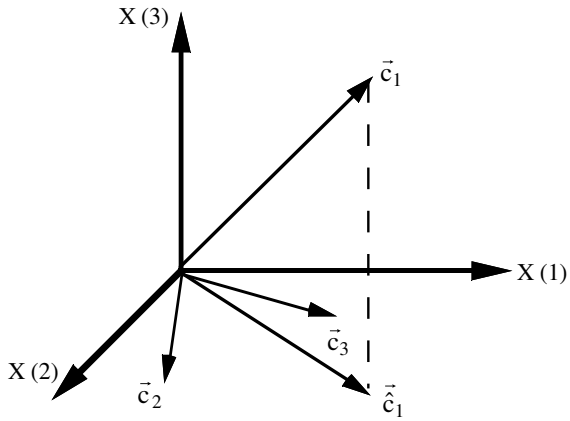
$$S^2_{c_1} = S^2_{\hat{c}_1} + S^2_e$$

edota

$$S^2_{x_1} = S^2_{\hat{x}_1} + S^2_e$$

Karratu txikienen erregresioa, aldagaien kopurua edozein izanik ere, grafikoki berdintsu interpreta daiteke. Aldagai dependentearen balio zentratuen bektorea erregresore guztiek osatzen duten azpiespazioan proiektatuz egingo da.

Hurrengo grafikoan, X_1 aldagaiaren erregresioa X_2, X_3 aldagaiekiko aurkezten dugu.



VIII. DOIKUNTZA ORTOGONALA ETA KOBARIANTZA MATRIZEAREN AUTODIREKZIOAK

VIII.1. SARRERA

VIII.2. ANALISIA $\mathbb{R}^2$ ESPAZIOAN. DOIKUNTZA ORTOGONALAREN ZUZENA

VIII.3. DOIKUNTZA ORTOGONALAREN ZUZENAREN LORPENA

VIII.4. HODEI DUALAREN AURKEZPEN GRAFIKOA

VIII.5. TRANSIZIO ERLAZIOAK

VIII.6. BATERAKO AURKEZPEN GRAFIKOA ETA INTERPRETAZIOA

VIII.7. DOIKUNTZA ORTOGONALAREN ZUZENA ALDAGAI NORMATUENTZAKO

VIII.1. SARRERA

Gai honen helburua, hurrengo gaian aurkeztuko dugun Osagai Nagusizko Analisiaren garapena ulergarriagoa egitea da.

Osagai Nagusizko Analisiaren helburua, berriz, dimentsioa murriztuz, (m x n) datu-matrizearen ikasketa burutzea da. Honek, aurkezpen grafiko batzuk, indibiduoen eta aldagaien interpretazioa eta berauen arteko erlazioak azaltzera eramango gaitu.

Demagun, adibidez, 100 automobil motaren ezaugarriei (motorra, leungailua, prezioa, garantia, konponketa zerbitzua...) dagozkien 20 aldagaien datuak jaso ditugula. Balio hauen (100 x 20) datu-matrizeak informazio aberatsa dauka baina automobilen arteko antzekotasunak eta ezberdintasunak ez dira erraz ikusiko.

Indibiduoak (automobilak) aldagai ezberdinetan hartzen dituzten balioen arabera parekatzea interesatzen zaigu, baina, aldagaien artean erlazioak egongo dira eta erlazio horietaz baliatuko gara daukaten dimentsioa murrizteko.

VIII.2. ANALISIA $\mathbb{R}^2$ ESPAZIOAN. DOIKUNTZA ORTOGONALAREN ZUZENA

Kasu honetan, bi aldagaien planotik bi ardatz nagusien planora iritsiko gara, lehen ardatzean hodeiaren informazio maximoa, posible den neurrian, proiektatzen delarik.

Lehen ardatz honi Doikuntza Ortogonalaren Zuzena (D.O.Z.) deituko zaio.

Jatorrizko datutatik abiatuz (x_{ij}), non i = 1, ..., m eta j = 1, 2, datu zentratu eta birreskalatuen matrizea osatuko dugu, horrela hurrengo eragiketa matematikoak erraztuko ditugu.

Honek (V.A.1.) ez ditu emaitzak aldatzen, egingo dugun analisia konparatiboa baita. Interesatzen zaiguna, aldagai ezberdinetan hartzen dituzten posizio erlatiboak aurkitzea da eta egindako transformazioak ez dauka eraginik.

Hau da, **X** datu-matrizea izango da:

$$\mathbf{X} = \left[\frac{x_{ij} - \bar{x}_j}{\sqrt{m}} \right]$$

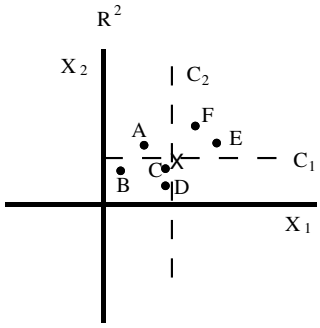
Gure helburua, puntu-hodeia dimentsio gutxi batzuetan proiektatzea da, honela, aurkezpen grafiko eskurakoiak edukiko ditugularik. Horretarako, metodo matematiko baten bidez, n aldagaien sistema batetik n ardatz berri dituen sistema batera iritsiko gara, hauek ordena beherakorra jarraituz, puntu-hodeiaren informazioa maximizatzen dutelarik.

Doikuntza Ortogonalaren kasuan, datu-taularen informazioa murrizteak ez dauka zentzu askorik, kasu honetan, metodoa bi aldagai bakarrik dituen kasurako konkretatzen baitugu. Honela, gai honen helburu bakarra, pedagogikoa da, eta n aldagaien kasurako egingo dugun analisi orokorra ulergarriagoa izango zaigu.

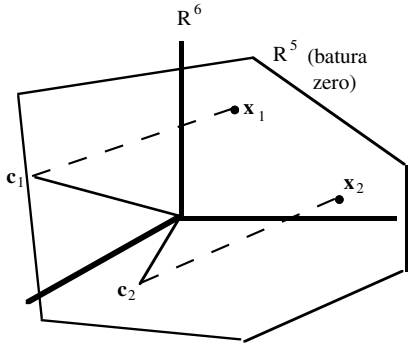
Hurrengo bost grafiken bidez, sei automobil mota direla suposatuz, prozesua laburtzen dugu, grafikoki.

JATORRIZKO ARDATZAK

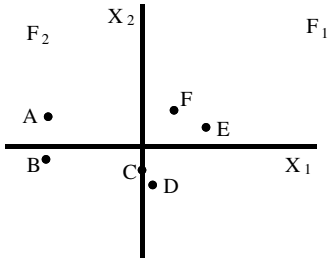
INDIBIDUOEN PUNTU-HODEIA

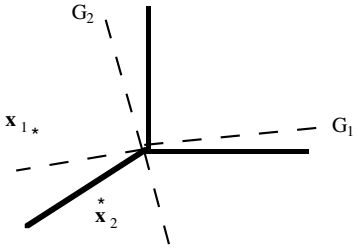


ALDAGAIEN PUNTU-HODEIA

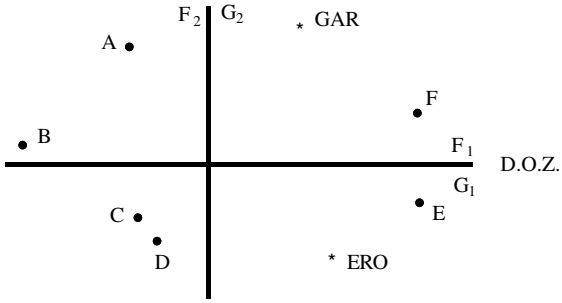


ARDATZ BERRIAK





BATERAKO AURKEZPEN GRAFIKOA



D.O.Z. lortu aurretik puntu-hodeiaren inertzia definituko dugu.

Definizioa: Puntu-hodei baten inertzia edozein punturekiko, erreferentzi puntu horretara, hodeiaren puntuen distantzi karratuen batura da.

Hodeia zentraturik dagoenez, inertzia jatorriarekiko edo grabitate-zentruarekiko kontsideratuko dugu. Hots:

$$I = \sum_{i=1}^{m} d^2(0, M_i) = \sum_{i=1}^{m} \left[\left(\frac{x_{i1} - \bar{x}_1}{\sqrt{m}} \right)^2 + \left(\frac{x_{i2} - \bar{x}_2}{\sqrt{m}} \right)^2 \right]$$

Aurrerantzean, $x_{i1} - \bar{x}_1 / \sqrt{m}$ eta $x_{i2} - \bar{x}_2 / \sqrt{m}$ balioak x_{i1} eta x_{i2} izango dira, hau da, balio zentratuak eta birreskalatuak.

Honela, M_i (x_{i1}, x_{i2}), i = 1,...,m puntu-hodeiaren puntuak izango dira.

Dakusagunez:

$$I = S_1^2 + S_2^2$$

Inertzia bi bariantzen batura da.

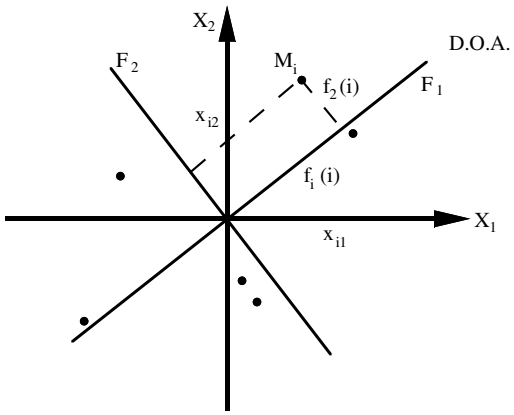
Inertzia, orduan, hodeiaren sakabanatze neurria da bere grabitate-zentruarekiko, eta puntu-hodeiaren informazioaz hitz egiten dugunean, datuen artean dagoen sakabanatzeaz ari gara.

F zuzen baten gainean edozein puntu proiektatzean, proiekzioaren f(i) balioa lortzen dugu. Gure helburua, hodeian hoberen doitzen den F_1 zuzena lortzea da;

hau da, puntu guztietatik hurbilen dagoena. Datuak zentraturik daudenez, zuzen horrek jatorritik iragan behar du, hau da, grabitate-zentrutik.

Hurrengo grafikoa ikusirik, ondoko erlazioa lortzen dugu:

$$d^2(0, M_i) = f_1^2(i) + f_2^2(i)$$



Honela, inertzia adieraz daiteke:

$$I = \sum_{i=1}^{m} \left[f_1^2(i) + f_2^2(i) \right]$$

non $f_1(i)$, i = 1,...,m, F zuzenaren gaineko proiekzioak diren; $f_2(i)$ balioak F_2 zuzenaren gaineko proiekzioen balio berdinak dira eta F_1 zuzenaren gainean proiektatzean dauzkagun erroreak edo hondarrak bezala interpreta daitezke.

Honela:

$$I_{F_1} = \sum_{i=1}^{m} f_1^2(i) \qquad I_{F_2} = \sum_{i=1}^{m} f_2^2(i)$$

Eta hodeiaren inertzia totala F_1 ardatzaren gain proiektatutako inertzian (I_{F_1}) eta F_2 ardatzaren gain proiektatutako inertzian (I_{F_2}) deskonposatzen da batukorki. Azken hori, hondar-inertzia bezala interpreta daiteke, F_1-ek bildu ez duen informazioa biltzen baitu.

Hodeiaren inertzia (I) konstantea denez, hodeian hoberen doitzen den zuzena, hondar-inertzia minimoa edo proiektatutako inertzia maximoa egiten duen zuzena izango da.

VIII.3. DOIKUNTZA ORTOGONALAREN ZUZENAREN LORPENA

Gure helburua proiektatutako inertziaren maximizazioa izango da.

Orokorki, zuzen baten unitate bektore zuzendaria **u** baldin bada, edozein **x** bektorearen proiekzioa zuzenaren gaineran $\mathbf{x}^T\mathbf{u}$ biderkadura eskalarraz lortuko da. Horrela, puntu-hodeiaren puntu guztien proiekzioak daukan bektorea **f**, hain zuzen, lortuko da:

$$\mathbf{f} = \mathbf{X}\mathbf{u}$$

Ebatzi behar dugun problema $\left\{\sum_{i=1}^{m} f^2(i)\right\}$ maximoa egitea da.

Aurreko erlazioa kontutan harturik, maximizazio problema adieraz daiteke:

$$\max_{\mathbf{u}} \ \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u}$$

$$\mathbf{u}^T\mathbf{u} = 1 \ \text{ baldintzaz}$$

forma koadratiko baldintzatu baten maximizazioaren aurrean aurkitzen gara eta lagrangiarra eratu behar dugu.

$$£ = \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} - \lambda\,(\mathbf{u}^T\mathbf{u} - 1) = \mathbf{u}^T\,\mathbf{L}\,\mathbf{u} - \lambda\,(\mathbf{u}^T\mathbf{u} - 1)$$

non, £-ren maximoa lortu behar dugun.

Deribazio matriziala (V.A.7.) aplikatuz, maximizazioaren baldintza beharrezkoak lortzen ditugu:

$$\frac{\partial £}{\partial \mathbf{u}} = 2\mathbf{L}\mathbf{u} - 2\lambda\mathbf{u} = \mathbf{0} \Rightarrow \mathbf{L}\mathbf{u} = \lambda\mathbf{u}$$

$$\frac{\partial £}{\partial \lambda} = -(\mathbf{u}^T\mathbf{u} - 1) = \mathbf{0} \Rightarrow \mathbf{u}^T\mathbf{u} = 1$$

Hau da, $\mathbf{u}_1$ soluzioa **L** kobariantza matrizearen autobektore unitarioa da. **L** matrize simetrikoa eta mugatu edo erdimugatu positiboa denez, $\lambda_1 \geq \lambda_2 \geq 0$ autobalio erreal eta positiboak daukala ziurtatuta dago.

Zein autobalioari dago lotuta $\mathbf{u}_1$ autobektore unitarioa?

Dakigunez $\mathbf{u}_1$-ek inertzia proiektatua maximizatu behar du, orduan:

$$\mathbf{u}_1^T \ \mathbf{L} \ \mathbf{u}_1 = \lambda \ \mathbf{u}_1^T\mathbf{u}_1 = \lambda = \lambda_1$$

Hots, **L**-ren λ_1 autobalioarik handiena izango da. Horrela F_1, D.O.Z.-ren direkzioa $\mathbf{u}_1$ autobektore unitarioaz finkatuta daukagu.

F2, **L**-ren bigarren autodirekzioa $\mathbf{u}_2$ autobektore unitarioaz finkatuko da. $\mathbf{u}_2$, $(\mathbf{u}_2^T\mathbf{u}_2 = 1$ eta $\mathbf{u}_2^T\mathbf{u}_1 = 0)$ unitarioa eta $\mathbf{u}_1$ autobektorearekin ortogonala izanik, λ_2 bigarren autobalioari dagokio.

F_1 eta F_2 ardatzek, ardatz-sistema berri bat osatzen dute.

Planoan daudenez D.O.Z. $\mathbf{u}_2$ bektorearekiko ortogonalak diren puntuen multzoaz adieraz daiteke. Izan bedi $\mathbf{u}_2 = (\alpha, \beta)^T$; zuzenaren adierazpena izango da:

$$\alpha \ \mathbf{X}_1 + \beta \ \mathbf{X}_2 = 0$$

$\mathbf{u}_2 = (\alpha, \beta)^T$ izanik F_1-en autobektore unitario bat $\mathbf{u}_1 = (-\beta, \alpha)^T$ izango da eta hodeiaren m puntuen proiekzioak edo koordenatuak zuzenaren gainean izango dira:

$$f_1(i) = -\beta \ x_{i1} + \alpha \ x_{i2}$$

matrizialki, $\mathbf{f}_1$ proiekzio guztien bektorea izanik

$$\mathbf{f}_1 = \mathbf{X} \ \mathbf{u}_1$$

Bestalde, puntu bakoitzaren errorea edo hondarra, hau da, $\mathbf{u}_2$ bektoreaz definitutako zuzenaren gaineko proiekzioa, horrela adieraziko da:

$$f_2(i) = \alpha \ x_{i1} + \beta \ x_{i2}$$

eta matrizialki, $\mathbf{f}_2$ proiekzioen bektorea izanik:

$$\mathbf{f}_2 = \mathrm{X} \ \mathbf{u}_2$$

Propietateak:

1. $\mathbf{f}_1$ eta $\mathbf{f}_2$ proiekzioen bektoreak aldagai zentratuak dira.

Biak aldagai zentratuen konbinazio linealak izanik, zuzenki ondorioztatzen da.

2. Proiekzioen bektoreak elkarren artean koerlazio gabeko aldagaiak dira.

$$\mathbf{f}_2^T\mathbf{f}_1 = \mathbf{u}_2 X^T X\mathbf{u}_1 = \mathbf{u}_2^T\lambda_1\mathbf{u}_1 = \lambda_1\mathbf{u}_2^T\mathbf{u}_1 = 0 \Rightarrow S_{F_2F_1} = 0$$

3. Ardatz bakoitzaren gainean proiektatutako inertzia ardatzari dagokion autobalioa da.

$$I_{F_1} = \sum_{i=1}^{m} f_1^2(c) = \mathbf{f}_1^T\mathbf{f}_1 = \mathbf{u}_1^T\mathbf{X}^T\mathbf{X}\mathbf{u}_1 = \mathbf{u}_1^T\lambda_1\mathbf{u}_1 = \lambda_1$$

Era berean $I_{F_2} = \lambda_2$

Eta honela idatz dezakegu:

$$\text{tr}(\mathbf{L}) = S_1^2 + S_2^2 = I = I_{F_1} + I_{F_2} = \lambda_1 + \lambda_2 = \text{tr}(\Lambda)$$

non, Λ matrizea, autobalioek osatzen duten matrize diagonala den.

Ardatzen fidegarritasuna. Ardatz bakoitzaren fidegarritasuna, beronen gain proiektatutako inertziaren arabera definituko dugu.

Puntu-hodeian D.O.Z. gero eta hobeago doitzen bada, inertzia totalarekiko proiektatutako inertziaren proportzioa gero eta handiagoa izango da. Doikuntza ezin hobea bada (puntuak lerrokaturik daude) inertzia guztia proiektatuta izango da eta hondar-inertziarik (edo bigarren ardatzan proiektaturik) ez daukagu.

Definizioa. τ fidegarritasunak, proiektatutako inertzia, inertzia totalarekiko bezala mugatzen du.

Hots:

$$\tau_1 = \frac{I_{F_1}}{I} = \frac{\lambda_1}{S_1^2 + S_2^2} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

$$\tau_2 = \frac{I_{F_2}}{I} = \frac{\lambda_2}{S_1^2 + S_2^2} = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

Ehunekotan emanda, bi ardatzeko fidegarritasunen batura % 100 izango da.

VIII.4. HODEI DUALAREN AURKEZPEN GRAFIKOA.

Indibiduoen puntu-hodeiaren analisian, lortu ditugun ardatzak aldagaien arabera interpretatzea gustatuko litzaiguke.

Azter dezagun aldagaien puntu-hodeia edo hodei duala, indibiduoen hodeiarekin egin dugun bezalaxe. Gure ikerketaren objektua hodei duala da, hau da, $\mathbf{X}^T$ matrizearen errenkadak, aldagai-puntuak $\mathbb{R}^m$-n dauden bi puntu dira (X_1 eta X_2). Dakigunez, $\sqrt{m}\ S_j$, aldagai zentratuaren norma da eta zentratuz gain aldagaia berreskalatzen badugu, aldagaiaren norma desbidazio tipikoarekin bat dator, horrela hodei dualaren inertzia:

$$I = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 = S_1^2 + S_2^2$$

Dakusagunez, indibiduo-puntuen inertzia berdina. Emaitza hori ez da harritzekoa, azken finean, bi hodeiak datu-matrize berberan daude eta informazio berbera daukate.

Berriro, hoberen doitzen den zuzena, G_1 izendatuko dugu, proiektatutako inertzia maximizatzen duena izango da. G edozein zuzenaren **v** unitate bektore zuzendaria izanik, aldagaien proiekzioak G-ren gainean bektorialki honela lortuko ditugu:

$$\mathbf{g} = \mathbf{X}^T \mathbf{v}$$

Maximizazio problema adieraz daiteke:

$$\max_{\mathbf{v}} \mathbf{v}^T\mathbf{X}\mathbf{X}^T \mathbf{v}$$

$$\mathbf{v}^T \mathbf{v} = 1 \quad \text{baldintzaz}$$

Lagrangiarra izango da:

$$£ = \mathbf{v}^T\mathbf{X}\mathbf{X}^T \mathbf{v} - \lambda (\mathbf{v}^T \mathbf{v} - 1)$$

Aurreko egoera berdinean gaude, orain $\mathbf{v} \in \mathbb{R}^m$ eta ikasketaren matrizea $\mathbf{X}\mathbf{X}^T$ da. $\mathbf{v}_1$ soluzioa, $\mathbf{X}\mathbf{X}^T$ matrizearen autobektore unitarioa, autobalio handienari dagokiona da. Autobektore horrek G_1 zuzena mugatzen du.

Dakigunez, λ, $\mathbf{X}^T\mathbf{X}$ matrizearen autobalioa denez:

$$\mathbf{X}^T\mathbf{X}\ \mathbf{u} = \lambda\ \mathbf{u} \Rightarrow \mathbf{X}\mathbf{X}^T\ (\mathbf{X}\mathbf{u}) = \lambda\ (\mathbf{X}\mathbf{u})$$

Hau da, $\mathbf{X}$ matrizeaz aurrebiderkatuz, λ, $\mathbf{XX}^T$ matrizearen autobalioa dela ikus dezakegu eta $\mathbf{XX}^T$, $\mathbf{X}^T\mathbf{X}$ matrize simetrikoak, ezberdin zero diren autobalio berberak dituzte. Autobalioak $\lambda_1 \geq \lambda_2 \geq 0$ izango dira; $\mathbf{XX}^T$ matrizeak, gainera, berdin zero diren m–2 autobalio ditu, bi matrizeen heina gehienez 2 baita, hots, $r(\mathbf{XX}^T) = r(\mathbf{X}^T\mathbf{X}) \leq 2$.

Hodei dualaren proiekzioak G_1 ardatzaren gainean izango dira:

$$\mathbf{g}_1 = \mathbf{X}^T \mathbf{v}_1$$

non, $\mathbf{v}_1$, $\mathbf{XX}^T$ matrizearen autobektore unitarioa, λ_1 autobalioari dagokiona den.

Izan bedi $\mathbf{v}_2$, λ_2 autobalioari dagokion autobektore unitarioa. G_2 ardatzaren gaineko proiekzioak izango dira:

$$\mathbf{g}_2 = \mathbf{X}^T \mathbf{v}_2$$

$\mathbf{XX}^T$ simetrikoa denez, lortutako ardatzak ortogonalak dira.

Propietateak:

1. **Aldagaien proiekzioen bektoreen biderkadura eskalarra zero da.**

$$\mathbf{g}_1^T\mathbf{g}_2 = \mathbf{v}_1^T\mathbf{XX}^T\mathbf{v}_2 = \mathbf{v}_1^T\lambda_2\mathbf{v}_2 = \lambda_2\mathbf{v}_1^T\mathbf{v}_2 = 0$$

2. **Ardatz bakoitzaren gainean proiektatutako inertzia ardatzari dagokion autobalioa da.**

$$I_{G_1} = g_1^2(1) + g_1^2(2) = \mathbf{g}_1^T\mathbf{g}_1 = \mathbf{v}_1^T\mathbf{XX}^T\mathbf{v}_1 = \mathbf{v}_1^T\lambda_1\mathbf{v}_1 = \lambda_1$$

$$I_{G_2} = g_2^2(1) + g_2^2(2) = \mathbf{g}_2^T\mathbf{g}_2 = \mathbf{v}_2^T\mathbf{XX}^T\mathbf{v}_2 = \mathbf{v}_2^T\lambda_2\mathbf{v}_2 = \lambda_2$$

Dakusagunez hodei dualaren ardatzen fidegarritasunak eta indibiduoen puntuhodeian lortu genituen ardatzenak berdinak dira. Hodei dualean, m–2 autodirekzio gehiago ditugu, baina, autobalio nuluei loturikoak direnez ez dute inertziarik jasotzen eta baztertu egingo ditugu.

VIII.5. TRANTSIZIO ERLAZIOAK

Ondoren, bi hodeien arteko proiekzioen eta bektore zuzendarien arteko erlazioak ikusiko ditugu, honela, ardatz berrien funtzioan indibiduoen eta aldagaien arteko erlazioak aztertu ahal izango ditugu.

$\mathbf{XX}^T$-ren λ_1 autobalioari dagokion autobektore unitarioa $\mathbf{v}_1$ denez:

$$\mathbf{XX}^T \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$$

$\mathbf{X}^T$ matrizeaz aurrebiderkatuz eta $\mathbf{g}_1 = \mathbf{X}^T \mathbf{v}_1$ dela jakinik, honako hau daukagu:

$$\mathbf{X}^T\mathbf{X} \mathbf{g}_1 = \lambda_1 \mathbf{g}_1$$

Eta dakusagunez, $\mathbf{g}_1$, $\mathbf{X}^T\mathbf{X}$ matrizearen λ_1 autobalioari lotutako autobektorea da, orduan, $\mathbf{u}_1$ autodirekzioan egongo den autobektore ez unitarioa, hain zuzen. $\|\mathbf{g}_1\| = \sqrt{\lambda_1}$ bere norma denez, idatz dezakegu:

$$\left.\begin{aligned} \mathbf{g}_1 &= \sqrt{\lambda_1}\mathbf{u}_1 \\ \mathbf{g}_2 &= \sqrt{\lambda_2}\mathbf{u}_2 \end{aligned}\right\} \text{ TRANTSIZIO ERLAZIOAK}$$

non, bigarren ekuazioko, $\mathbf{g}_2$ eta $\mathbf{u}_2$-ren arteko erlazioa, analogikoki, idatzi dugun.

Era berean, $\mathbf{X}^T\mathbf{X}$-en λ_1 autobalioari dagokion autobektore unitarioa $\mathbf{u}_1$ denez:

$$\mathbf{X}^T\mathbf{X} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

$\mathbf{X}$ matrizeaz aurrebiderkatuz eta $\mathbf{f}_1 = \mathrm{X}\mathbf{u}_1$ dela jakinik, honako hau daukagu:

$$\mathbf{XX}^T \mathbf{f}_1 = \lambda_1 \mathbf{f}_1$$

Eta dakusagunez, $\mathbf{f}_1$, $\mathbf{XX}^T$ matrizearen λ_1 autobalioari lotutako autobektorea da, orduan,$\mathbf{v}_1$ autodirekzioan egongo den autobektore ez unitarioa, hain zuzen. $\|\mathbf{f}_1\| = \sqrt{\lambda_1}$ bere norma denez, idatz dezakegu:

$$\left.\begin{aligned} \mathbf{f}_1 &= \sqrt{\lambda_1}\mathrm{v}_1 \\ \mathbf{f}_2 &= \sqrt{\lambda_2}\mathrm{v}_2 \end{aligned}\right\} \text{ TRANTSIZIO ERLAZIOAK}$$

non, bigarren ekuazioko, $\mathbf{f}_2$ eta $\mathbf{v}_2$-ren arteko erlazioa, analogikoki, idatzi dugun.

Laburki esanda, $\mathbb{R}^2$ edo $\mathbb{R}^m$ espazioko autobektoreak, $\mathbb{R}^m$ edo $\mathbb{R}^2$ bere hodei dualeko puntuen proiekzioen bektoreekin elkartzen dira.

$\mathbf{u}_1 = (-\beta, \alpha)^T$ eta $\mathbf{u}_2 = (\alpha, \beta)^T$ direla jakinik, aldagaien proiekzioak, honela lortuko ditugu:

$$\mathbf{g}_1 = \sqrt{\lambda_1}\begin{pmatrix} -\beta \\ \alpha \end{pmatrix}; \qquad \mathbf{g}_2 = \sqrt{\lambda_2}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

VIII.6. BATERAKO AURKEZPEN GRAFIKOA ETA INTERPRETAZIOA

Bi hodeiak aztertzean lortutako emaitzak, esangura berbera daukate, nahiz eta, indibiduoen funtzioan batean eta aldagaien funtzioan bestean adieraziak izan.

Adibidez, $\mathbf{f}_1$ indibiduoen proiekzioen bektorea kontsideratuz, trantsizio erlaziotik lortzen dugu:

$$\mathbf{f}_1 = \mathbf{X}\mathbf{u}_1 = \frac{1}{\sqrt{\lambda_1}}\mathbf{X}\mathbf{g}_1$$

Hau da:

$$\begin{pmatrix} f_1(1) \\ \vdots \\ f_1(i) \\ \vdots \\ f_1(m) \end{pmatrix} = \frac{1}{\sqrt{\lambda_1}}\begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{i1} & x_{i2} \\ \vdots & \vdots \\ x_{m1} & x_{m2} \end{pmatrix}\begin{pmatrix} g_1(1) \\ g_1(2) \end{pmatrix}$$

Orduan:

$$\mathbf{f}_1(i) = \frac{1}{\sqrt{\lambda_1}}\left(x_{i1}g_1(1) + x_{i2}g_1(2)\right)$$

Dakusagunez, indibiduo baten proiekzioa F_1-en gainean, G_1-en gaineko aldagaien proiekzioen konbinazio lineala da. Konbinazio linealaren koefizienteak indibiduo bakoitzarako aldagaien balio zentratuak dira. Indibiduoak aldagai baterako hartzen duen balioa, batezbestekoa baino handiagoa denean, koefiziente positiboa da eta txikiagoa denean, negatiboa. Hau da, grafiko berean indibiduoen eta aldagaien proiekzioak adieraziz, honako hau daukagu: indibiduo bakoitza, batezbestekoa baino balio handiagoak hartzen dituen aldagaien alde berean proiektatuko da. Hauxe da, baterako aurkezpen grafikoaren arrazoia.

F_1, G_1 eta F_2, G_2 izaeraz ezberdinak badira ere (F_1-en bektore zuzendaria $\mathbf{u}_1 \in \mathbb{R}^2$ eta G_1-en bektore zuzendaria $\mathbf{v}_1 \in \mathbb{R}^m$) batera aurkeztuko ditugu aldagaiek indibiduoak eta indibiduoek aldagaiak interpretatzera laguntzen baitute. Indibiduoen arteko distantziak ikusiz beraien arteko antzekotasunak eta ezberdintasunak ikus daitezke; halaber, aldagaiekin batera, badakigu grafikoki ere adieraztean zeintzuk diren antzekotasun edo ezberdintasun horien kausa.

VIII.7. DOIKUNTZA ORTOGONALAREN ZUZENA ALDAGAI NORMATUENTZAKO

Aldagaien eskala ezberdina izanik, analisian eragina eduki ez dezan, analisia aldagai normatutarako egingo da. Kasu honetan datu-matrize normatuaz abiatuz doikuntza ortogonalaren zuzenarako adierazpen orokorra daukagu, halaber, hondar-ardatzaren edo bigarren ardatzaren adierazpena, eta ondorioz, indibiduo- eta aldagai-puntuen proiekziotarako.

Kasu honetan:

$$\mathbf{X} = \left(\frac{x_{ij} - \bar{x}_j}{S_j \sqrt{m}} \right) \qquad \text{non} \quad i = 1,\ldots,m \quad \text{eta} \quad j = 1,2$$

$\mathbf{X}^T\mathbf{X}$ diagonalizatzen den matrizea, koerlazio-matrizea da eta bi aldagai besterik ez ditugunez, $|\mathbf{R} - \lambda I| = 0$ ekuazio karakteristikoa, honako hau da:

$$\begin{vmatrix} 1-\lambda & r \\ r & 1-\lambda \end{vmatrix} = 0$$

Ekuazioa ebatziz, λ-ren bi balioak: (1+r) eta (1–r) dira.

Lehen kasua: $r > 0$

Autobaliorik handiena $\lambda_1 = 1+r$ izango da eta $\lambda_2 = 1-r$ txikiena.

α eta β balioak lortzeko $\mathbf{R}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2$ ebatziko dugu:

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1-r) \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Hortik $\alpha = -\beta$. Autobektoreak unitarioa izan behar duenez:

$$\alpha = 1/\sqrt{2} = 0.707 \quad \text{eta} \quad \beta = -1/\sqrt{2} = -0.707$$

Honela, D.O.Z.-ren adierazpen orokorra:

$$0.707\ \mathbf{X}_1 - 0.707\ \mathbf{X}_2 = 0$$

lehen eta hirugarren koadranteen erdikariari dagokion zuzena, hain zuzen ere.

Autobektore unitarioak, honako hauek dira:

$$\mathbf{u}_1 = \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix} \qquad \mathbf{u}_2 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}$$

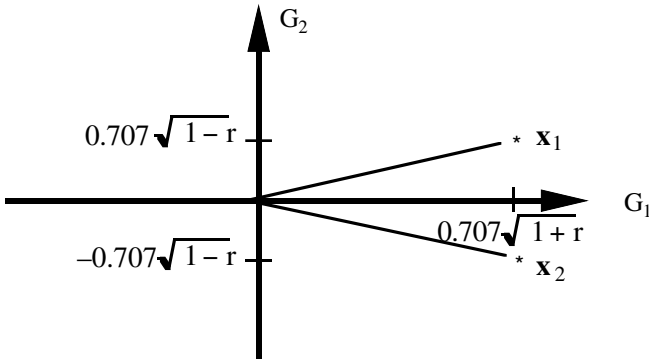
Eta indibiduoen (i=1,...,m) ardatzen gaineko proiekzioak:

$$f_1(i) = 0.707\ x_{i1} + 0707\ x_{i2}$$
$$f_2(i) = 0.707\ x_{i1} - 0707\ x_{i2}$$

Aldagaien proiekzioak berriz:

$$g_1(1) = 0.707\sqrt{1+r} \qquad g_2(1) = 0.707\sqrt{1-r}$$

$$g_1(2) = 0.707\sqrt{1+r} \qquad g_2(2) = -0.707\sqrt{1-r}$$



Analitikoki nahiz grafikoki ikus dezakegunez, aldagaiak G_1 lehen ardatzarekiko simetrikoak dira. r ~ 1 baldin bada $g_1(j) \sim 1$ izango da eta aldagai puntuak G_1-gaineko proiekzioetatik oso hurbil egongo dira; eta r = 1 baldin bada batera etorriko dira. r ~ 0 baldin bada bi ardatzetan aldagai bakoitzaren proiekzioak berdintsuak izango dira eta r = 0 baldin bada berdinak. Ondoren, $g_k(j)$ proiekzioa j. aldagaiaren eta k. ardatzaren arteko koerlazioa dela ikusiko dugu.

Trantsizio erlaziotik lortzen dugu:

$$\mathbf{g}_k = \mathbf{X}^T \frac{\mathbf{f}_k}{\sqrt{\lambda_k}} \qquad k = 1,2$$

Orduan:

$$g_k(j) = \mathbf{x}_j^T \frac{\mathbf{f}_k}{\sqrt{\lambda_k}} = r_{\mathbf{x}_j,\mathbf{f}_k} \qquad k = 1,2 \quad \text{eta} \quad j = 1,2$$

Bigarren kasua: $r < 0$.

Kasu honetan, ondoko emaitzak lortzen ditugu:

$$\lambda_1 = 1 - r \quad , \quad \lambda_2 = 1 + r$$

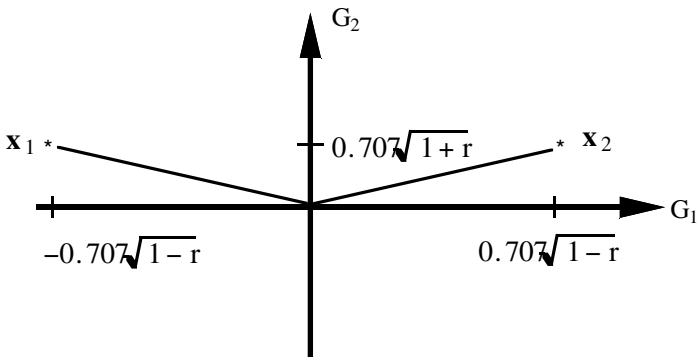
D.O.Z. $0.707\ \mathbf{X}_1 + 0.707\ \mathbf{X}_2 = 0$ da; indibiduoen proiekzioak:

$$f_1(i) = -0.707\ x_{i1} + 0707\ x_{i2}$$
$$f_2(i) = \ 0.707\ x_{i1} + 0707\ x_{i2}$$

eta aldagaien proiekzioak:

$$g_1(1) = -0.707\sqrt{1-r} \qquad g_2(1) = 0.707\sqrt{1+r}$$

$$g_1(2) = 0.707\sqrt{1-r} \qquad g_2(2) = 0.707\sqrt{1+r}$$



Kasu honetan, G_2 bigarren ardatzarekiko aldagaien posizioa simetrikoa da. Koerlazioa ($r \sim -1$) –1etik hurbil baldin badago $g_1(j)$ proiekzioen tamainua 1etik hurbil egongo da eta ($r \sim 0$) koerlazio ahula baldin badago proiekzioen tamainua

0tik hurbilago egongo da konkretuki r = 0 baldin bada bi ardatzen gaineko proiekzioak aldagai bakoitzarako, balio absolututan, berdinak izango dira.

Bi kasuetan, hau da, $r > 0$ edo $r < 0$ denean aldagaien arteko koerlazioa gero eta indartsuagoa denean, gero eta errepresentagarriagoa, izango da lehen ardatza. Lehen kasuan ($r > 0$), lehen ardatzak bi aldagaietan balio altuak (batezbestekoa baino handiagoak) hartzen dituzten indibiduoak eta balio baxuak (batezbestekoa baino txikiagoak) hartzen dituztenak kontrajartzen ditu. Hau gertatzen denean, "neurri" efektuaren aurrean gaudela esaten da, ardatzak indibiduoak neurtzen baititu. Bigarren kasuan ($r < 0$) alde berean dauden inbibiduek, alde horretan kokatzen den aldagairako balio altuak eta beste aldagairako balio baxuak hartzen dituzte.

Lehen ardatzaren fidegarritasuna $\tau_1 = (1 + |r|)/2$ da eta bigarren ardatzarena $\tau_2 = (1 - |r|)/2$.

IX. DATU ANIZKOITZEN ANALISIA

IX.1. SARRERA

IX.2. ANALISI OROKORRA

IX.2.1. $\mathbb{R}^n$ espazioaren $\mathbb{R}^q$ azpiespazio baten bidez egindako doikuntza

IX.2.2. $\mathbb{R}^m$ espazioaren $\mathbb{R}^q$ azpiespazio baten bidez egindako doikuntza

IX.2.3. $\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioen arteko erlazioa

IX.2.4. X datu taularen berreraketa

IX.3. OSAGAI NAGUSIZKO ANALISIA

IX.3.1. $\mathbb{R}^n$ espazioan egindako analisia

IX.3.2. $\mathbb{R}^m$ espazioan egindako analisia

IX.3.3. Baterako aurkezpen grafikoak

IX.3.4. Adibidea

IX.4. KORRESPONDENTZI ANALISI FAKTORIALA

IX.4.1. Hodeiak, masak eta distantziak

IX.4.2. $\mathbb{R}^n$ espazioan egindako analisia

IX.4.3. $\mathbb{R}^m$ espazioan egindako analisia

IX.4.4. $\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioen arteko erlazioa

IX.4.5. Maiztasun-taularen berreraketa

IX.4.6. Interpretaziorako laguntzak

IX.4.7. Korrespondentzia Anitzeko Analisi Faktoriala

IX.5. ELEMENTU GEHIGARRIAK

IX.6. SAILKAPEN-METODOAK

IX.6.1. Goranzko Sailkapen Hierarkikoa

IX.6.1.1. Bariantzaren metodoa

IX.6.2. Adibidea

IX.1. SARRERA

Azken hogei edo hogeita hamar urteetan, konputagailu eta kalkulu automatikoarekin batera, estatistika deskribatzaile anizkoitza asko garatu da J.P. Benzecri-ren eskola frantsesan.

Datu-Analisiaz ezaguten diren metodo hauek, bi motakoak dira:

- Metodo faktorialak
- Sailkapen-metodoak

Datu-multzo handi berdinei aplikatzen zaizkie, berorien ikasketa egiterakoan osagarri bezala erabiltzen direlarik.

Liburu honen lehen gaiaren sarreran esaten genuenez, estatistika deskribatzailearen helburua datu-multzo handien ikasketak egitea da, haien barne egitura gardenduz.

Ikasketa horietan, metodo faktorialak eta sailkapen-metodoak, datu-taula handien analisi bakoitzean bata besteari lagunduz, dira urrunago eramaten gaituztenak.

Zenbakizko datuen bidez, errealitate anizkoitzaz partzialki ohartu ahal badugu ere, metodo estatistiko egokiak aplikatzen dituzten kalkulu-programen bidez, errealitatean dauden estrukturak aztertu ahal ditugu.

Laburki esanez, metodo faktorialak diagonalizazio-metodoak dira, hauen bidez datutan dauden ardatz nagusi batzu eskurakoi bihurtzen direlarik.

Normalki, deskribatu nahi ditugun objetuak aurkezpen grafiko jarraietan kokatzen ditugu; ondoren, taldekatzen saiatzen gara puntu-hodeian dauden konstelazioak ikusi ahal izateko, eta sailkapen metodoen bidez eskurakoi bihurtzen dira.

Sailkapen metodoak, metodo algoritmikoak dira; egokia den algoritmo bat era errepikariaz aplikatuz oharpenen edo indibiduoen sailak edo klaseak lortzen ditugu.

Gai honetan, metodo faktorialen oinarrian dagoen Analisi orokorra aurkezten da, ondoren, Osagai Nagusizko Metodoa eta Korrespondentzi Analisi

Faktoriala ikusiko dira, ondoren Korrespondentzia Anitzeko Analisi Faktorialaren ideia aurkezten da eta azkenik, sailkapen metodoaz arituz Bariantzaren Metodoa ere laburki aurkeztuko da.

IX.2. ANALISI OROKORRA

Atal honetan, analisi faktorialen doikuntza metodoek duten gune teorikoa azalduko dugu.

Edozein metrika edo Analisi Kanonikoa erabiltzeak aurkezpen dotoreagoa egiteko aukera emango liguke, baina oso utilak ez diren garapen matematikotan sartuz, Datu-Analisiaren oinarrian dagoen ideia nagusia ilunduta utziko genuke ikasleentzat.

Ideia hauxe da: Datu-matrize batek, bi espazio bektorial desberdinetan, hodei itxurako aurkezpen grafikoak egiteko aukera ematen du eta hodei horietan egindako doikuntzak, erlazio sinple batzuen bidez loturik daude.

Izan bedi $\mathbf{X}(m,n)$ datu-matrizea. Bere elementu orokorraz adieraziz $\mathbf{X} = [x_{ij}]$ izango da eta m.n balioez osaturik dago.

m.n balio kopurua oso handia izan daiteke eta honen aurrean gure galdera hauxe da: aurkez ditzakegu m.n balioak balio kopuru txiki baten bidez, taulan daukagun sakabanatze adieragarrienaz jabetuz?

Demagun $\mathbf{u}_1$ eta $\mathbf{v}_1$, hurrenez hurren, m eta n osagai dituzten zutabe--bektoreak direla eta berorien bidez $\mathbf{X}$ datu-matrizea (m+n) balio horien bidez berreratu ahal dugula.

$$\underset{(\mathbf{m},\mathbf{n})}{\mathbf{X}} = \underset{(m,1)}{\mathbf{u}_1}\ \underset{(1,n)}{\mathbf{v}_1^T}$$

Kasu honetan $\mathbf{X}$ datu-matrizearen heina 1 da. Normalki, oso zaila izango da $\mathbf{X}$ horrela berreratzea, orduan, q heinaren hurbilketa bilatuko da.

Hau da:

$$\mathbf{X} = \mathbf{u}_1 \mathbf{v}_1^T + \mathbf{u}_2 \mathbf{v}_2^T + \ldots + \mathbf{u}_q \mathbf{v}_q^T + \mathbf{E}$$

$\mathbf{E}(m,n)$ hondar-matrizea izanik.

$\mathbf{u}_k$ $\mathbf{v}_k$ bektoreen q(m + n) osagaien bidez **X** matrizearen m.n osagaiak, modu egokian berreratu edo berreraiki daitezke, **E** matrizearen osagaiak txikiak eta baztergarriak baitira.

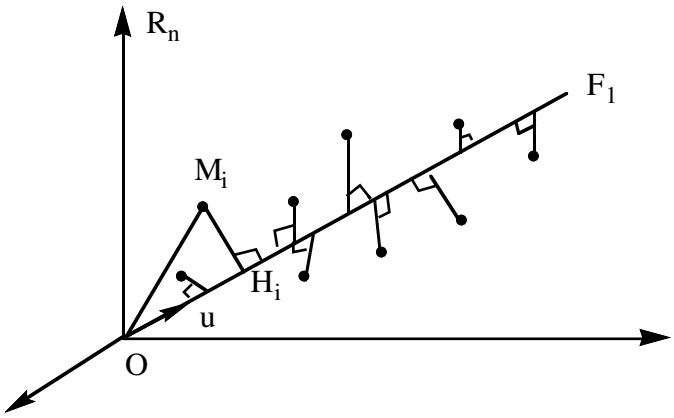
Aurreneko problema, metodo faktorialei loturiko aurkezpen geometrikoen bidez ebatziko dugu:

X datu-taulatik, dakigunez (ikus X.2) bi aurkezpen grafiko eratortzen dira, m errenkadak, $\mathbb{R}^n$ espazio bektorialeko m punturen koordenatuak bezala kontsidera daitezke eta n zutabeak, $\mathbb{R}^m$ espazio bektorialeko n punturen koordenatuak bezala. Hau da, errenkada-puntuen hodeia eta zutabe-puntuen hodeia $\mathbb{R}^n$ eta $\mathbb{R}^m$ espaziotan ditugularik.

IX.2.1. $\mathbb{R}^n$ espazioaren $\mathbb{R}^q$ azpiespazio baten bidez egindako doikuntza

Aurrerago ikusiko dugunez, $\mathbb{R}^n$ espazioan ditugun m errenkada-puntuak, $\mathbb{R}^q$ azpiespazio bateko ardatzetan dituzten koordenatuen bidez eta ardatz berri hauek dituzten osagaien bidez, berrera daitezke. Honela, taularen m.n balioak q(m + n) = qm + qn balioen bidez berreratuak geratuko dira.

Orduan, ohizko distantzia euklidearraz hornitutako $\mathbb{R}^n$ espazioko $\mathbb{R}^q$ azpiespazio baten bidez hodeia doitzea da gure helburua. Hasieran, hoberen doituz, jatorritik pasatzen den F_1 zuzena bilatuko dugu:



Izan bedi **u** zuzen horren bektore unitarioa. Orduan,

$$\mathbf{u}^T\mathbf{u} = 1 \quad \text{eta} \quad \sum_{i=1}^{n} \mathbf{u}_i^2 = 1$$

X datu-taularen errenkadetan $\mathbb{R}^n$ espazioko m puntuak ditugunez **u** bektore unitarioaz atzebiderkatuz F_1 zuzenean puntu horien proiekzioak edo koordenatuak lortuko ditugu

Hau da: $\underset{(\mathbf{m,n})}{\mathbf{X}} \quad \underset{(n,1)}{\mathbf{u}} = \underset{(m,1)}{\mathbf{f}}$

f bektorearen m osagaiak, F_1 zuzenaren gain lortzen dituzten luzerak dira.

$\mathbb{R}^2$-espazioan planteatzen genuen doikuntza ortogonalaren gaian gertatzen zen bezalaxe (ikus VIII.2), hemen ere, $\mathbb{R}^n$ espazioan, hain zuzen, puntu bakoitzetik jatorrira dagoen distantzia karratua bi distantzia karratutan deskonposatzen da, hauek, F_1 zuzenera daukan e_i distantzia karratua eta F_1 gainean daukan f_i proiekzioaren karratua direlarik.

Irudian: $\overline{OM_i}^2 = \overline{MH_i}^2 + \overline{OH_i}^2$ edo $\overline{OM_i}^2 = e_i^2 + f_i^2$

Ondorioz, puntuek ardatzera daukaten distantzia karratuak minimizatzea (karratu txikienen erizpidea) edo zuzenaren gain daukaten proiekzio karratuak maximizatzea batera ematen da.

Puntu guztiak batera hartuz:

$$\sum_i \overline{OM_i}^2 = \sum_i \overline{MH_i}^2 + \sum_i \overline{OH_i}^2$$

$\sum_i e_i^2$ minimoa edo $\sum_i f_i^2$ maximoa bilatzea baliokidea da.

Arazoa, bada, honela aurkez daiteke:

Zein da, proiekzio $\mathbf{f}^T\mathbf{f} = \sum_i f_i^2$ karratuen batukaria maximoa egiten duen F_1 ardatzaren $\mathbf{u}_1$ bektore unitarioa?.

Hots: $\mathbf{f}^T\mathbf{f} = \mathbf{u}_1^T\mathbf{X}^T\mathbf{X}\ \mathbf{u}_1$ maximoa $\mathbf{u}_1^T\mathbf{u}_1 = 1$ baldintzaz.

Honela, hodeia hoberen doitzen duen dimentsio bateko azpiespazioa lortzen da, hau da, $\mathbf{u}_1$-ek sortarazten duen azpiespazioa, $\mathbf{u}_1^T\mathbf{X}^T\mathbf{X}\mathbf{u}_1$ forma koadratikoa maximoa egiten duelarik.

Hodeia hoberen doitzen duen bi dimentsioko azpiespazioak $\mathbf{u}_1$ daukala, absurdora eramanez erraz froga daiteke: $\mathbf{u}_1$ ez balu, $\mathbf{u}_1$ edukiz hobe doituko lukeen bi dimentsioko beste azpiespazioa lortuko genuke.

Honela, $\mathbf{u}_2$ bektore unitarioa eta $\mathbf{u}_1$-ekiko ortogonala $(\mathbf{u}_2^T \mathbf{u}_2 = 1$ eta $\mathbf{u}_2^T \mathbf{u}_1 = 0)$ lortuz, $\mathbf{u}_2^T \mathbf{X}^T \mathbf{X}\, \mathbf{u}_2$ forma koadratikoa maximoa egiten duen bi dimentsioko azpiespazioa sortaraziko da.

Analogikoki, era errepikariaz arrazonatuz zera ikusiko dugu: $\mathbf{u}_1$, $\mathbf{u}_2$,...$\mathbf{u}_q$ bektore unitarioek, non $\mathbf{u}_q$, aldez aurretik lortutako $\mathbf{u}_1$,$\mathbf{u}_2$,...,$\mathbf{u}_{q-1}$ bektoreekiko ortogonala baita, $q \leq n$ dimentsioko azpiespazioa sortarazten dute, $\mathbf{u}_q^T \mathbf{u}_q = 1$ baldintzaz, $\mathbf{u}_q^T \mathbf{X}^T \mathbf{X}\, \mathbf{u}_q$ forma koadratikoa maximoa egiten delarik.

Maximo baldintzatuaren ebazpidea: Izan bedi λ Lagrange-ren biderkatzailea, $\mathbf{u}^T\mathbf{u} = 1$ izanik, $\mathbf{u}^T\ \mathbf{X}^T\ \mathbf{X}\ \mathbf{u} - \lambda\ (\mathbf{u}^T\ \mathbf{u} - 1)$ adierazpenaren maximoa lortu behar dugu.

Dakigunez, optimizazio-baldintza beharrezkoak aplikatuz, $\mathbf{u}$ bektorearen osagai bakoitzarekiko deribatuak zero egin behar dira.

Osagai bakoitzarekiko deribatuak eginez, zero balioari berdinduz, eta gero ditugun ekuazioak matrizialki idatziz ondoko ekuazio matriziala daukagu: (berdin litzateke matrizialki deribatzea)

$$2\ \mathbf{X}^T\ \mathbf{X}\ \mathbf{u} - 2\ \lambda\ \mathbf{u} = 0$$

Hau da: $$\mathbf{X}^T\ \mathbf{X}\ \mathbf{u} = \lambda\ \mathbf{u} \qquad (1)$$

Dakusagunez, λ balioa eta $\mathbf{u}$ bektorea $\mathbf{X}^T\ \mathbf{X}$ matrize simetrikoaren autobalioa eta berari loturiko autobektorea, hurrenez hurren, dira.
(1) berdintasuna, $\mathbf{u}^T$ bektoreaz aurrebiderkatuz:

$$\mathbf{u}^T\ \mathbf{X}^T\ \mathbf{X}\ \mathbf{u} = \lambda$$

Hau da, bilatzen ari garen maximoa λ autobalioa da. Honela, $\mathbf{u}$ bektorea λ_1 autobalio handienari dagokion $\mathbf{X}^T\ \mathbf{X}$ matrizearen $\mathbf{u}_1$ autobektorea izango da.

Hoberen doitzen den bi dimentsioko azpiespazioa lortu nahi badugu, bigarren zuzen bat behar dugu $\mathbf{t}$ berorren bektore unitarioa izanik.

$\mathbf{t}^T\ \mathbf{X}^T\mathbf{X}\,\mathbf{t}$ maximoa $\mathbf{t}^T\mathbf{u}_1 = 0$ eta $\mathbf{t}^T\mathbf{t} = 1$ izanik lortu behar dugu.

Hau da: $\mathbf{t}^T\ \mathbf{X}^T\mathbf{X}\,\mathbf{t} - \lambda\,(\mathbf{t}^T\mathbf{t} - 1) - \mu\mathbf{t}^T\mathbf{u}_1$, bi biderkatzaile dituen lagrangearra idazten dugu.

Berriro ere, $\mathbf{t}$ bektorearen osagaiekiko deribatu partzialak zero balioari berdinduz, ondoko ekuazio matriziala daukagu.

$$2\,\mathbf{X}^T\mathbf{X}\,\mathbf{t} - 2\,\lambda\,\mathbf{t} - \mu\,\mathbf{u}_1 = 0 \qquad (2)$$

(2) berdintasuna $\mathbf{u}_1^T$ bektoreaz aurrebiderkatuz, ondoren $\mathbf{u}_1^T\mathbf{X}^T\mathbf{X} = \lambda\mathbf{u}_1^T$ ordezkatuz eta $\mathbf{u}_1^T\ \mathbf{t} = 0$, $\mathbf{u}_1^T\ \mathbf{u}_1 = 1$ direla jakinik $\mu = 0$ dela ikusten dugu.

Hau da: $\mathbf{X}^T\mathbf{X}\,\mathbf{t} = \lambda\,\mathbf{t}$

Dakusagunez, $\mathbf{t}$ bektore unitarioa, $\mathbf{X}^T\mathbf{X}$ matrize simetrikoaren λ_2 bigarren autobalioari loturiko $\mathbf{u}_2$ bigarren autobektorea izango da. Prozesua jarraituz, k = 1,2....,r autoelementuetara edo elementu propioetara zabaltzen da emaitza.

Honela, $\mathbf{X}^T\mathbf{X}\,\mathbf{u}_k = \lambda_k\ \mathbf{u}_k \qquad \forall\, k \le r \qquad (3)$

r, $\mathbf{X}^T\mathbf{X}$ matrizearen heina izanik.

Azkenik, karratu txikienen zentzuan hodeian hoberen doitzen den q dimentsioko azpiespazioaren oinarria, $\mathbf{X}^T\mathbf{X}$ matrize simetrikoaren q handienen autobalioei loturiko q autobektoreez osaturik dago.

$\mathbf{X}^T\mathbf{X}$ matrizea, simetrikoa eta definitu edo erdidefinitu positiboa izanez, hauek dituzten propietateetatik abiaturik, beste frogapen baten bidez, aurreneko emaitzara iritsiko ginateke.

IX.2.2 $\mathbb{R}^m$ espazioaren $\mathbb{R}^q$ azpiespazio baten bidez egindako doikuntza

Era berean, $\mathbb{R}^m$ espazioan ditugun n zutabe-puntuak $\mathbb{R}^q$ azpiespazioko ardatzetan dituzten koordenatuen bidez eta ardatz berri hauek dituzten osagaien bidez berrera daitezke.

Izan bedi G_1 jatorritik pasatzen den zuzena eta $\mathbf{v}$ beronen direkzio--kosinuen bektorea; G_1 zuzenaren gainean puntuen proiekzioen karratuen batura maximoa egitean, hodeian hoberen doitzen den zuzena, karratu txikienen zentzuan lortzen dugu.

Hau da: $\mathbf{X}^T\ \mathbf{v} = \mathbf{g}$
$(n.m)(m,1)\ \ (n,1)$

g zutabe-bektorearen n errenkadak proiekzioen n balioak dira, orduan, $\mathbf{v}^T\ \mathbf{X}\ \mathbf{X}^T\ \mathbf{v}$ forma koadratikoa, maximoa egin nahi dugun balioa, izango da.

$\mathbf{v}^T\ \mathbf{v} = 1$ baldintzaz, $\mathbf{v}^T\mathbf{X}\ \mathbf{X}^T\mathbf{v}$ maximoa egiten duen **v** bektorea aurkituz, problema ebatzia edukiko dugu.

$\mathbb{R}^n$ espazioaren kasuan bezala, matrize simetriko baten diagonalizazioaren aurrean gaude, $\mathbb{R}^m$ espazioan, $\mathbf{X}\mathbf{X}^T$ matrize simetrikoaren autobektorea **v** baita.

μ_1 autobalio handienari dagokion autobektorea $\mathbf{v}_1$ izango da. Ondoren, $\mathbf{v}_2$ ... $\mathbf{v}_q$ finkatuz, hodeian hoberen doitzen den q dimentsioko azpiespazioa sortarazten duten bektoreak lortu ditugu.

Dakigunez, $\mathbf{X}^T\ \mathbf{X}$ eta $\mathbf{X}\ \mathbf{X}^T$ matrizeak **X** matrizearen heina berbera daukate r, hain zuzen, eta $q < r$ izango da.

IX.2.3 $\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioen arteko erlazioa

$\mathbf{v}_k$ eta $\mathbf{u}_k$ bektoreen arteko erlazioak aurkitu nahi ditugu.

Definizioz $\mathbf{X}\ \mathbf{X}^T\ \mathbf{v}_k = \mu_k\ \mathbf{v}_k$ (4)

μ_k eta $\mathbf{v}_k$, $\mathbf{X}\ \mathbf{X}^T$ matrizearen k. autobalioa eta berari loturiko k. autobektorea, hurrenen hurren, izanik.

Diagonalizazioaren bidez, desberdin zero diren r autobalio lortuko dira eta $r \leq (m,n)$, hau da, **X** matrizearen dimentsiorik txikiena baino txikiagoa edo berdina izango da r, **X** matrizearen heina r delako.
(4) berdintasuna $\mathbf{X}^T$ matrizeaz aurrebiderkatuz:

$$\mathbf{X}^T\ \mathbf{X}\ (\ \mathbf{X}^T\ \mathbf{v}_k\ ) = \mu_k\ (\ \mathbf{X}^T\ \mathbf{v}_k\ )$$

Baina, $\mathbf{X}^T\ \mathbf{X}$ diagonalizazioan $\mathbf{X}^T\ \mathbf{X}\ \mathbf{u}_k = \lambda_k\ \mathbf{u}_k$ ((3) berdintasuna) daukagunez, $\mathbf{X}\ \mathbf{X}^T$ matrizearen $\mathbf{v}_k$ bakoitzari ($k \leq r$) $\mathbf{X}^T\ \mathbf{X}$ matrizearen $\mathbf{X}^T\ \mathbf{v}_k$ dagokio μ_k autobalio berdinari lotuta.

Dakusagunez, $\mathbf{u}_k = h_k \mathbf{X}^T \mathbf{v}_k$ (h_k = konstantea) idatz dezakegu bektore proportzionalak baitira.

λ_k k. autobalioa, handiena izatean, nahitaez:
$\mu_k \leq \lambda_k$ izango da, eta $\{\mu_k\} \subset \{\lambda_k\}$

Hots, μ_k autobalioen multzoa λ_k autobalioen multzoaren azpimultzoa da.

Era berean, (3) berdintasuna **X** matrizeaz aurrebiderkatuz:

$$\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{u}_k) = \lambda_k(\mathbf{X}\mathbf{u}_k)$$

Honela, $\mathbf{X}^T\mathbf{X}$ matrizearen $\mathbf{u}_k$ bakoitzari $\mathbf{X}\mathbf{X}^T$ matrizearen $\mathbf{X}\mathbf{u}_k$ dagokio, λ_k autobalio berdinari lotuta.

Orduan, $\mathbf{v}_k = h_{k'}\mathbf{X}\mathbf{u}_k$ (h_k konstantea) eta μ_k, $\mathbf{X}\mathbf{X}^T$ matrizearen k. autobalio handiena denez

$\lambda_k \leq \mu_k$ eta $\{\lambda_k\} \subset \{\mu_k\}$

Hots: $\lambda_k = \mu_k \quad \forall k \leq r$ idatz dezakegu.

Dakigunez, $\mathbf{u}_k^T\mathbf{u}_k = 1, \quad \mathbf{v}_k^T\mathbf{v}_k = 1$

Honela, $\mathbf{v}_k^T\mathbf{v}_k = h_k^2, \quad \mathbf{u}_k^T\mathbf{X}^T\mathbf{X}\mathbf{u}_k$

eta $\mathbf{X}^T\mathbf{X}\mathbf{u}_k = \lambda_k\mathbf{u}_k$ ordezkatuz.

$$\mathbf{v}_k^T\mathbf{v}_k = h_k^2, \mathbf{u}_k^T\lambda_k\mathbf{u}_k = h_k^2\lambda_k = 1$$

$$h_{k'} = \frac{1}{\sqrt{\lambda_k}}$$

Analogikoki, $h_k = \frac{1}{\sqrt{\lambda_k}}$ dela lortzen da.

Orduan,

$$\underset{(n,1)}{\mathbf{u}_k} = \frac{1}{\sqrt{\lambda_k}} \underset{(n,m)(m,1)}{\mathbf{X}^T \mathbf{v}_k} \qquad (5)$$

$$\underset{(m,1)}{\mathbf{v}_k} = \frac{1}{\sqrt{\lambda_k}} \underset{(m,n)(n,1)}{\mathbf{X}\,\mathbf{u}_k} \qquad (6)$$

$\mathbf{u}_k$ bektore unitarioaren euskarria den F_k ardatza, $\mathbb{R}^n$ espazio bektorialaren k. ardatz faktoriala deitzen da eta $\mathbf{v}_k$ bektore unitarioaren euskarria den G_k ardatza $\mathbb{R}^m$ espazio bektorialaren k. ardatz faktoriala, hain zuzen ere, biek, λ_k autobalio berdinari dagokion autodirekzioa finkatzen dutelarik.

Dakusagunez, $\mathbb{R}^n$ (edo $\mathbb{R}^m$)-ren k.ardatzaren gain hodeiaren puntuek dituzten koordenatuak (osagai nagusiak) $\mathbf{X}\,\mathbf{u}_k$ (edo $\mathbf{X}^T\,\mathbf{v}_k$) dira. Orduan, (5), (6) erlazioetan ikus daitekeenez, koordenatu horiek espazio batean eta ardatzaren osagai unitarioak beste espazioan proportzionalak dira.

IX.2.4 X datu taularen berreraketa

Dakusagun, nola, $\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioetan ditugun hodeien puntuak beraien posizioetan, gutxi gora-behera, berreratzen ditugun karratu txikienen zentzuan hoberen doitzen den azpiespazio baten oinarrian dituzten koordenatuen bidez.

Suposa dezagun, λ_1, lehen autobalioa, besteak baino askoz garrantzitsuagoa dela. Kasu honetan, $\lambda_1 = \mathbf{u}_1^T \mathbf{X}^T \mathbf{X}\,\mathbf{u}_1$ balioak proiekzioz jatorriarekiko distantzia karratuak neurtzen ditu eta lehen ardatzak puntuen posizioak ongi berreratuko ditu. Normalki, q lehen ardatz faktorialen bidez, "hurbilketa onaz", puntuen posizioak berreratuko ditugu, horretarako $\lambda_1 + \lambda_2 + ... + \lambda_q$ baturak, $\mathbf{X}^T\,\mathbf{X}$ matrizearen aztarnaren proportzio garrantzitsua aurkeztu behar du. $\mathbf{X}^T\,\mathbf{X}$ matrize simetrikoa eta, normalki, definitu positiboa izango denaren $n = r \leq m$ autobalioak errealak dira eta autobektoreak ortogonalak.

U autobektore unitarioak osatutako matrize diagonalizatzailea eta $\mathbf{X}^T\,\mathbf{X}$ matrizea antzekoak dira, hau da, oinarri desberdinetan transformazio lineal berbera definitzen dute.

Hots, $\mathbf{\Lambda}$ matrizea, $\lambda_1, \lambda_2 \ldots \lambda_n$ autobalioak osaturiko $\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$ matrize diagonala izanez, $\mathbf{X}^T\,\mathbf{X}\,\mathbf{U} = \mathbf{U}\,\mathbf{\Lambda}$ eta $\mathbf{U}^T\,\mathbf{X}^T\,\mathbf{X}\,\mathbf{U} = \mathbf{U}^T\,\mathbf{U}\,\mathbf{\Lambda}$ edota $\mathbf{U}^{-1}\,\mathbf{X}^T\,\mathbf{X}\,\mathbf{U} = \mathbf{\Lambda}$, $\mathbf{U}$ ortogonala baita.

Dakusagunez, bada, $\mathbf{X}^T\,\mathbf{X}$ matrizearen antzekotasunez transformatu diagonala da $\mathbf{\Lambda}$ matrizea eta honelako matrize antzekoen propietate bat aplikatuz[1]:

$$\mathrm{tr}(\mathbf{X}^T\mathbf{X}) = \mathrm{tr}\ \mathbf{\Lambda}$$

$$\mathrm{tr}(\mathbf{X}^T\mathbf{X}) = \sum_i \sum_j x_{ij}^2 = \sum_k \lambda_k = \mathrm{tr}\ \mathbf{\Lambda} \qquad (7)$$

Hasierako zenbakizko balioak, hau da, $\mathbb{R}^n$ espazioan puntuak dituzten koordenatuak, q ardatzen bidez lortu nahi izanez gero, ardatz faktorialetan dituzten koordenatuak eta ardatzen osagai unitarioak (direkzio-kosinuak) batera erabiliz lortuko ditugu.

(6) erlazioa honela idaziz: $\mathbf{X}\,\mathbf{u}_k = \sqrt{\lambda_k}\ \mathbf{v}_k$

Ondoren, berdintasuna $\mathbf{u}_k^T$ bektoreaz atzebiderkatuz:

$$\mathbf{X}\,\mathbf{u}_k\,\mathbf{u}_k^T = \sqrt{\lambda_k}\,\mathbf{v}_k\,\mathbf{u}_k^T$$

eta k balio guztientzako batukaria hartuz (autobalio nuluak balira beraiei dagozkien $\mathbf{u}_k$ autobektoreek $\mathbb{R}^n$ espazioaren oinarria osatuko lukete),

$$\mathbf{X} \sum_{k=1}^{n} \mathbf{u}_k\,\mathbf{u}_k^T = \sum_{k=1}^{n} \sqrt{\lambda_k}\ \mathbf{v}_k\ \mathbf{u}_k^T$$

eta $\underset{(mn,n)}{\mathbf{X}} = \sum_{k=1}^{n} \sqrt{\lambda_k}\ \underset{(m,1)(1,n)}{\mathbf{v}_k\,\mathbf{u}_k^T} = \mathbf{u}\mathbf{u}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$ (unitate matrizea)

q lehen ardatzak hartzen baditugu, lortzen dugun $\mathbf{X}^*$ matrizea $\mathbf{X}$ datu-matrizearen hurbilketa izango da.

1. Lau propietate dituzte, diskriminatzailea, autobalioak, aztarrena (tr) eta heina (r) berbera edukitzea, hain zuzen. Ikus FZ. DE TROCONIZ A. *Matemáticas Generales* I, 1.18.2

$$X \simeq X^* = \sum_{k=1}^{q} \sqrt{\lambda_k}\ \mathbf{v}_k\, \mathbf{u}_k^T$$

$\sqrt{\lambda_{q+1}} \ldots \sqrt{\lambda_n}$ balioak txikiak balira hurbilketa ona izango litzateke.

Berreraketaren kalitatea, ondoko proportzioaz neur daiteke:

$$\tau_q = \sum_{k=1}^{q} \lambda_k \Big/ \sum_{k=1}^{n} \lambda_k$$

X datu-matrizearen m.n balioak q (m+n) = q m + q n balioen bidez berreratu ditugu $\sqrt{\lambda_k}\ \mathbf{v}_k$, q bektoreen m balioen bidez eta $\mathbf{u}_k$, q bektoreen n balioen bidez.

λ_k autobalio bakoitzak k. ardatzaren gain puntuen proiekzio karratuen batura neurtzen du, hau da, k. ardatzaren gain proiektatutako hodeiaren inertzia (ikus VIII.3 n=2 kasurako). Honela, inertzi-tasa deitutako τ_q balioak q lehen ardatzetan proiektatutako inertziaren proportzioa edo bariantzaren proportzioa ematen dizkigu.

Hoberen doitutako $\mathbb{R}^q$ azpiespazioan biltzen den hodeiaren sakabanatzearen proporzioa, inertzi-tasak neurtzen digu.

IX.3. OSAGAI NAGUSIZKO ANALISIA

Osagai nagusizko analisia, aztertu dugun analisi orokorraren barrutian aurkezten dugu.

Osagai nagusizko analisia "aldagai-indibiduo" taula motatan aplikatuko da, kasu honetan, errenkada-puntuen hodeia, indibiduoen hodeia eta zutabe-puntuen hodeia aldagaien hodeia izango direlarik.

Datu zentratutatik abiaturik eta hodei batean nahiz bestean zentratze--eragiketak suposatzen duena kontutan hartuz, osagai nagusizko analisia bi eratara plantea daiteke.

Osagai Nagusizko Analisia

X matrizearen gai orokorra $\dfrac{x_{ij} - \bar{x}_j}{\sqrt{m}}$ denean, diagonalizatzen den matrize simetrikoa $\mathbf{X}^T\mathbf{X} = \mathbf{L}$ kobariantza matrizea da.

Osagai Nagusizko Analisi Normatua.

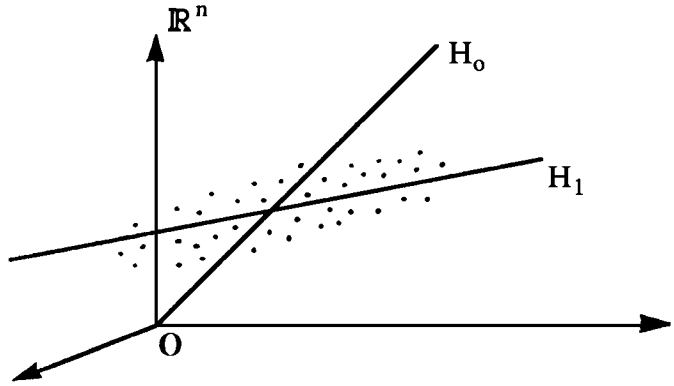
Aldagaiaren neurriak, batezbestekoen aldetik eta sakabanatzearen aldetik heterogenoak direnean **X** matrizearen gai orokorra $\frac{x_{ij} - x_j}{S_j \sqrt{m}}$ izango da, kasu honetan diagonalizatzen den matrize simetrikoa $\mathbf{X}^T \mathbf{X} = \mathbf{R}$ koerlazio-matrizea da.

IX.3.1 $\mathbb{R}^n$ espazioan egindako analisia

Datu zentratutatik abiatzeak dituen abantailak hauexek dira:

- Diagonalizatzen den matrizea **L** edo **R** da.
- Proiekzio karratuen batura maximoa egiten duen lehen ardatza jatorritik pasatzen da; honela aztertu dugun analisi orokorraren barrutian azaldu ahal dugu osagai nagusizko analisia.

Ondoko grafikoan ikus dezakegunez, hodeiaren grabitate-zentrua jatorrian ez balego, proiekzio karratuen batura maximoa egingo lukeen ardatza ez litzateke jatorritik pasatuko.



H_1 ardatzarentzako balio hori maximoa izango litzateke eta ez H_o ardatzarentzako.

$\mathbb{R}^n$ espazioan, indibiduoen hodeian egindako analisian, grabitate-zentrua jatorrira eramaten da edota ardatzen translazio paraleloa egiten da. $1/\sqrt{m}$ koefizientearen bidez, diagonalizatzen den $\mathbf{X}^T \mathbf{X}$ matrizea kobariantza matrizea izatea lortzen da.

Aldagaien sakabanatzeak oso desberdinak direnean, zentratzeaz gainera, beste aldaketa egitea beharrezkoa da, honela, osagai normatutako analisian indibiduoen arteko distantziak aldagai desberdinekiko, ondoan ikusten den bezala "eskalan berdindu ditugu".

Hots:

$$d^2(i,i') = \sum_{j=1}^{n} \left[\frac{x_{ij} - \bar{x}_j}{S_j \sqrt{m}} - \frac{x_{i'j} - \bar{x}_j}{S_j \sqrt{m}} \right]^2 =$$

$$= \frac{1}{m} \sum (x_{ij} - x_{i'j})^2 / S_j^2$$

Dakusagunez, indibiduoen hurbiltasunak kalkulatzean, aldagaiek ekarpen analogoa daukate, desbidazio tipikoen arabera hartu baititugu koordenatuak.

Kasu honetan, dakigunez, diagonalizatzen den $\mathbf{X}^T \mathbf{X}$ matrizea koerlazio--matrizea da.

Indibiduo-puntuek k. ardatzaren gain dituzten koordenatuak, $\mathbf{Xu}_k$ zutabe--bektorearen errenkadak dira, $\mathbf{Xu}_k = \sqrt{\lambda_k}\ v_k$ (6) formularen arabera.

IX.3.2 $\mathbb{R}^m$ espazioan egindako analisia

Espazio honetan ditugun n puntuak aldagai-puntuak dira. Dakigunez, $\mathbb{R}^n$ espazioan egindako analisiak $\mathbb{R}^m$ espazioaren analisia induzitzen du. Halere, i eta j azpiindizeek betebehar desberdinak dituzte eta **X** datu-matrizean egindako transformazioek interpretazio geometriko desberdinak dituzte (ikus VII. gaia).

$\mathbb{R}^m$ espazioan, aldagaien hodeian, zentratze-eragiketaren bidez aldagai--puntuak $\mathbb{R}^{m-1}$ azpiespazio batera proiektatzen ditugu, konkretuki, lehen erdikariarekiko proiekzio paraleloa eginez, koordenatuen batura zero ematen duen $\mathbb{R}^{m-1}$ azpiespaziora proiektatzen ditugu.

Aldagai-puntuak jatorrira daukan distantzia karratua aldagaiaren bariantza da:

$$d^2(j,o) = \frac{1}{m} \sum_{i=1}^{m} \left(x_{ij} - \bar{x}_j \right)^2 = S_j^2$$

j,j' bi aldagaien arteko distantzia karratua:

$$d^2(j,j') = \frac{1}{m}\sum_{i-1}^{m}\left[\left(x_{ij} - \bar{x}_j\right) - \left(x_{ij'} - \bar{x}_{j'}\right)\right]^2$$

$$= S_j^2 + S_{j'}^2 - 2S_{jj'}$$

Aldiz, Osagai normatutako analisian ardatzen eskalak aldatzeak, hau da, $\mathbb{R}^m$ espazioan ditugun koordenatuak, $S_i\sqrt{m}$ aldagaiaren normaren balioaz zatituz, aldagai-puntuak normatzen ditugu eta guztiak jatorriarekiko 1 distantziaz kokatzen dira.

Hots:

$$d^2(j,o) = \frac{1}{m}\sum_{i=1}\left(x_{ij} - \bar{x}_j\right)^2 / S_j^2 = 1$$

n aldagai-puntuak, koordenatuen jatorrian zentraturiko erradioz bat daukan hiperesfera batean kokatzen dira.

Ondokoak, j eta j' bi aldagai puntuen arteko distantzia karratua:

$$d^2(j,j') = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{x_{ij} - \bar{x}}{S_j} - \frac{x_{ij'} - \bar{x}_{j'}}{S_{j'}}\right]^2$$

Karratua eginez eta batukaria aplikatuz, hauxe da ateratzen dugun emaitza:

$$d^2(j,j') = S_j^2 / S_j^2 + S_{j'}^2 / S_{j'}^2 - 2r_{jj'} = 2(1 - r_{jj'})$$

Honela, aldagai-puntuen hurbiltasunak koerlazioen arabera interpreta daitezke. Oso hurbil dauden bi aldagai-puntuen arteko koerlazioa positiboa eta sakona izango da ($r_{jj'} \simeq 1$), eta oso urrun dauden bi aldagai-puntuen koerlazioa negatiboa eta sakona izango da ($r_{jj'} \simeq -1$). Aldagai-puntuek k. ardatzaren gain dituzten koordenatuak $\mathbf{X}^T\mathbf{v}_k$ zutabe-bektorearen osagaiak dira, eta (5) formularen arabera $\mathbf{X}^T\mathbf{v}_k = \sqrt{\lambda_k}\,\mathbf{u}_k$.

Dakusagunez, bada, aldagai-puntuen koordenatuak k. ardatzaren gain, $\mathbf{R}$ koerlazio-matrizearen diagonalizazioan kalkulatutako λ_k, $\mathbf{u}_k$, kantitateen funtzioan kalkulatzen dira.

Dakigunez, (ikus VII. gaia) bi aldagai-puntuk osatzen duten angeluaren kosinua aldagaien arteko koerlazio-koefizientea da. Analisi normatuaren kasuan, aldagai-puntuak zentru-normatuak direnez, kosinua aldagaien biderkadura eskalarra da.

Aldagai-puntuek k. ardatzaren gain dituzten abzisak $\mathbf{X}^T \mathbf{v}_k$ zutabe--bektorearen osagaiak dira.

Dakusagunez, abzisa bakoitza aldagai-bektore bakoitzaren ($\mathbf{X}^T$ matrizearen errenkada) eta $\mathbf{v}_k$ bektorearen arteko biderkadura eskalarra da. Honela, aldagai--puntu bakoitzak ardatz baten gain daukan abzisa, aldagaia eta aldagai guztien konbinazio lineala den $\mathbf{v}_k$ aldagai berriaren arteko koerlazio-koefizientea da.

Analisi Orokorrean ikusten genuenez, λ_k autobalio bakoitzak k. ardatzaren gain proiekzio karratuen batura neurtzen du, hau da, k. ardatzaren gain proiektatutako hodeiaren inertzia.

$\mathbf{X}^T \mathbf{X}$ matrizeak, inertzi-matrizea izanik, bere aztarrenean hodeiaren jatorriarekiko inertzia biltzen du.

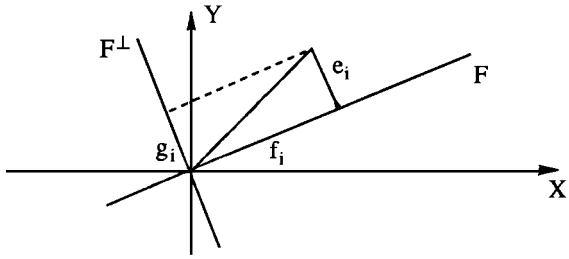
Bestalde, diagonalizazioan egindako transformazio lineal ortogonalaren bidez, jatorriarekiko baterako inertzia, n ardatz ortogonaletan proiektatutako inertzitan deskonposatzen da (ikus (7) berdintasuna).

$\mathbb{R}^n$, n-dimentsioko espazio bektoriala, dimentsio bateko n azpiespazio ortogonaletan deskonposatu da, n ardatz faktorialetan, hain zuzen.

$\mathbb{R}^n$, n–dimentsioko espazio bektoriala, bi espazio ortogonaletan deskonposa dezakegu, hauek betegarriak direlarik.

Honela, q lehen ardatzetan hodeiaren sakabanatzearen zatirik adierazgarriena biltzen denean, $\mathbb{R}^n$ espazioa $\mathbb{R}^q$ eta $\mathbb{R}^{n-q}$ azpiespazio betegarritan deskonposatzen da.

n = 2 denean (ikus VIII gaia) oso erraz ikus daiteke hodeiaren baterako inertzia F zuzen baten gainean proiektatutako inertzian eta zuzen horrekiko daukan hondar-inertzian deskonposatzen dela, konkretuki, F zuzenaren ortogonala den $F^{\perp}$ zuzenaren gainean proiektatutako inertzia eta F-rekiko hondar-inertzia berdinak dira.



Puntu bakoitzerako $g_i = e_i$ eta puntu guztientzako

$$\sum_i g_i^2 = \sum_i e_i^2$$

hodeiaren baterako inertzia batukorki deskonposatzea, Pitagoras-en teoremaren aplikatzeaz zuzenki eratortzen delarik.

$n > 2$ denean eta $\mathbb{R}^n$ espazioa $\mathbb{R}^q$ eta $\mathbb{R}^{n-q}$ azpiespazio betegarritan deskonposatuta daukagunean, hodeiaren inertzia bi espaziotan proiektatutako inertzien batura da, baina $\mathbb{R}^{n-q}$ espazioan proiektatutako inertzia eta $\mathbb{R}^q$ espazioarekiko hondar-inertzia berdinak dira.

Era berean, hodeiaren baterako inertzia batukorki deskonposatzea, "Pitagoras-en Teorema Espazioan" aplikatzeaz zuzenki eratortzen da.

Osagai Nagusiko Analisian, Doikuntza Ortogonalean gertatzen zen bezala, diagonalizatzen dugun matrizea **L** kobariantza matrizea edo **R** aldagai tipifikatuen kobariantza matrizea da. Halere, $\left(x_{ij} - \bar{x}_j\right)\sqrt{m}$ edo $\left(x_{ij} - \bar{x}_j\right) S_j \sqrt{m}$ aldagai transformatuen inertzia matrizeak bezala lortu direla gogoratu behar dugu.

Edozein kasutan, λ_k autobalio bakoitzak k. ardatzaren gain puntuen proiekzioen karratuen batura batezbestekoz neurtzen du, hau da, k. ardatzaren gain proiektatutako hodeiaren bariantza.

Hodeiaren baterako bariantza:

$$\text{tr } \mathbf{L} = \sum_{j=1}^{n} S_j^2 = \sum_{k=1}^{n} \lambda_k = \text{tr } \Lambda$$

edo

$$\text{tr } \mathbf{R} = n = \sum_{k=1}^{n} \lambda_k = \text{tr } \Lambda$$

kasu honetan, koerlazio-matrizearen elementu diagonalak 1 baitira.

Era berean, q lehen ardatzetan hodeia ondo doituta kontsideratzen badugu, $\sum_{k=1}^{q}\lambda_k$ proiektatutako bariantza eta $\sum_{k=q+1}^{n}\lambda_k$ hondar-bariantza izango dira.

Azkenik:

$$\tau_q = \frac{\sum_{k=1}^{q}\lambda_q}{\sum_{k=1}^{n}\lambda_k}$$

bariantza tasa izango da, proportzio horrek, askotan portzentaiatan emanik, q lehen faktoreetan proiektatutako bariantzaren portzentaia neurtzen du, edo q lehen faktoreetan proiektatutako inertziaren portzentaia, honela, era berean, inertzia tasa deritzogu.

IX.3.3. Baterako aurkezpen grafikoak

Osagai nagusiko analisia mende honen hasieratik ezaguna da, baina, azken hogeita hamar urte hauetan eduki duen bultzada oso handia izan da.

Arlo askotan aplikaturik, Soziologian, Ekonomian, Biologian, Geologian, Antropologian, Linguistikan... datu-taula anizkoitzetatik laburpen deskribatzaileak ateratzeko guztiz erabilgarria da.

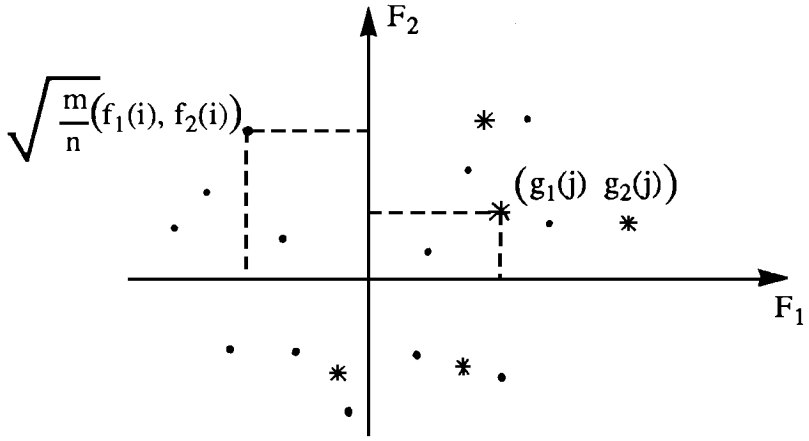
Analisi-teknika honen garapena, gaiaren hasieran esaten genuen bezala, konputagailuen eta kalkulu automatikoen garapenari loturik dago, baina praktikan, erabilpenaren garapena, baterako aurkezpen grafikoetatik eratorritako interpretagarritasunean datza.

$\mathbf{X}^T\mathbf{X}$ eta $\mathbf{X}\mathbf{X}^T$ matrize simetrikoen analisietan lortutako emaitza paraleloek, hau da, autobalio berdinak eta autodirekzio baliokideak edukitzeak, errenkada--puntuak eta zutabe-puntuak hodeiaren ebaketa-plano berean proiektatzearen ideia sortu zuten Datu-Analisiaren eskola frantsesean.

Trantsizio-erlazioak deitzen diren (5) eta (6) erlazioetan daukagu indibiduo--puntuak eta aldagai-puntuak erlazionatzeko bidea.

(6) erlazioa ikusiz zera esan dezakegu: k. ardatzaren gain koordenatu handienak dauzkaten indibiduo-puntuek koefiziente handiagoz hartzen dute parte aldagai-puntuen hodeian k. ardatza finkatzean. Eta, (5) erlazioa ikusiz, hodeia eta puntu-mota aldatuz, ondorio berdina atera daiteke.

Bi ardatz faktorialak osatutako planoa $\mathbb{R}^n$ edo $\mathbb{R}^m$ espazioetan, jatorritik pasatuz egindako ebaketa da, eta lehen plano faktoriala edo plano faktorial nagusia deritzogu.



Analisiaren praktikan beste plano batzu ere aurkezten dira (F_1, F_3) (F_2, F_3)..., alegia, kasu bakoitzerako estatistikariak adierazgarriagoak kontsideratzen dituenak.

Analisi normatua egin bada, aldagaien koordenatuak bat baino txikiagoak izango dira, erradioz bat daukan hiperesfera batean daudela gogoratu behar dugu eta proiekzio-eragiketa murriztzaile dela. $\mathbb{R}^m$ espazioan ebaketa egitean, erradioz bat daukan zirkulua ebaki da eta puntuak zirkunferentziatik hurbil eduki ahal ditugu.

Dakusagunez, indibiduoen hodeian, analisia grabitate-zentruarekiko egiten da, baina ez aldagaien hodeian, horrela aldagai-puntu denak planoaren alde berean egon daitezke; aski da horretarako beraien arteko koerlazio guztiak positiboak izatea.

Bi hodeien inertziak berdinak izatean, eta normalki aldagaien kopurua indibiduoen kopurua baino askoz txikiago izatean, lehenengoak planoan proiektatzean jatorritik urrunago geratuko dira. Konkretuki, ardatz bakoitzean proiektatzean indibiduo-puntuaren distantzia tipikoa jatorriarekiko $\sqrt{\lambda_k/m}$ izango da, eta $\sqrt{\lambda_k/n}$ aldagai-puntuarena. Orduan, indibiduo-puntuen koordenatuak $\sqrt{m/n}$ balioaz biderkatzen dira aldagai-puntuekin planoan bateragarriak izateko.

Ardatzen alde positiboa eta negatiboa edonolakoa da eta diagonalizazioaren algoritmoaren menpe izango da. $\mathbb{R}^n$ espazioan k. ardatzaren bektore unitarioa aukeratzeko, bi posibilitate daude, $\mathbf{u}_k$ edo - $\mathbf{u}_k$, hain zuzen, baina bata aukeratuz gero $\mathbb{R}^m$ espazioarena finkatuta dago, dakigunez $\mathbf{v}_k = h_k$; $\mathbf{X}\,\mathbf{u}_k$ eta honen ondorioz (5), (6) trantsizio-erlazioak espazio baten elementuen abzisak k. ardatzaren gain, beste espazioan, k. ardatzaren bektore unitarioaren osagaien bidez ematen dizkigutenak, koherenteki aplikatutako dira.

Bi hodeien puntuak plano faktorialetan proiektatu ondoren eta adierazpide grafiko hauetatik ondorioak atera baino lehen, ondokoa kontutan hartu behar dugu:

Aldagai batzuren posizioek, inguruan dauden indibiduoen hurbiltasunaren interpretazioaren ikuspegia emango digute, baina aldagai bat indibiduo baten ondoan egoteak ez du esangurarik, puntu baten kokapena beste hodeiaren puntu guztien kokapenarekin dago erlazionaturik eta hortik datorkio bere posizioa. Ez dugu inoiz ahaztu behar ikuspegi estatistikoa multzokoa delarik.

IX.3.4. Adibidea

Aurreko garapen teorikoen argibide gisa erabil daitekeen adibidea hauxe da:

Suposa dezagun Hego eta Ertamerikako ikasketa bat egin nahiean, hamazortzi estaturako sei aldagaien ohartutako balioak ditugula.

Aukeratu ditugun ondoko aldagaietan, munduko kontinente honen errealitatea partzialki ohartzen dugu; halere, errealitateari hurbiltzen zaion adibide honetan, ikasleak SPAD paquete-programaren azpiprograma batzuren bidez lortutako balioak jarrai ditzake kalkuladora baten bidez.

Aldagaiak

Datu-matrizeko zutabeetan ditugun aldagaiak eta identifikatzaileak hauek dira:

Nazio-Produktu Gordina / biztanle, dolarretan: ..... NPGB
Urteroko inflazio-tasa, 1980-1985 tarterako: ..... UINT
Biztanle kopurua, milaka biztanletan, 1985 urterako: .... BIZK
Populazioaren hazkunde-tasa, 1973-1985 tarterako: ..... POHT
Hiri-populazioaren ehunekoa, 1985 urterako: ..... HIPO
Eskolako matrikula-tasa, 1984 urterako: ..... ESMT

Datu-matrizea:

Estatuak	**Ident.**						
VENEZUELA	VEN	3110	9.2	17323	3.3	85	74
ARGENTINA	ARG	2130	342,8	30531	1.6	84	88
MEXIKO	MEX	2080	62.2	78820	2.8	69	88
PANAMA	PAN	2020	3.7	2180	2.3	50	83
URUGUAI	URU	1660	44.6	3004	0.6	85	86
BRASIL	BRA	1640	147.7	135539	2.3	73	78
TXILE	TXI	1440	19.3	11990	1.7	83	95
KOLONBIA	KOL	1320	22.5	24418	1.9	67	73
COSTA RICA	CRI	1290	36.4	2593	2.8	45	69
GUATEMALA	GUA	1240	7.4	7966	2.8	41	47
EKUADOR	EKU	1160	29.7	9367	2.9	52	85
PERU	PER	960	98.6	18653	2.4	68	85
PARAGUAI	PAR	940	15.8	3388	2.5	41	73
NIKARAGUA	NIK	850	33.8	3263	3.1	56	71
DOMINIKAR ERREPUBLIKA	DOE	810	14.6	6261	2.4	56	81
HONDURAS	HON	730	5.4	4366	3.5	39	67
EL SALVADOR	ELS	710	11.6	5564	3.0	43	56
BOLIVIA	BOL	470	569.1	6383	2.7	44	71

Argiro ikusten denez, aldagaiak oso heterogenoak dira eta honek Osagai Nagusizko Analisi Normatua egitea eskatzen du.

Programaren bidez jasotako lehen estatistiko bakunetan, batezbestekoen eta desbidazio tipikoen desberdintasunak ikus daitezke, halaber, berorien bidez atera ditzakegun aldakuntza koefizienteen desberdintasunak.

Estatistiko bakunak

1 DESCRIPTION SOMMAIRE DES VARIABLES SUR 18 INDIVIDUS

	VARIABLE	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM	ABSENTS
1 /	NPGB	1364.4446	639.2376	470.0000	3110.0000	0
2 /	UINT	81.9111	142.0133	3.7000	569.1000	0
3 /	BIZK	20644.9414	32960.8437	2180.0000	°°°°°°°°°°	0
4 /	POHT	2.4778	.6762	.6000	3.5000	0
5 /	HIPO	60.0556	16.4096	39.0000	85.0000	0
6 /	ESMT	76.1111	11.6232	47.0000	95.0000	0

Ondoren jasotzen dugun koerlazio-matrizea oso adierazgarria da, adibide txiki honetan, sei aldagai besterik ez baititugu.

Koerlazio-matrizea:

	NPGB	UINT	BIZK	POHT	HIPO	ESMT
NPGB	1.00	-.14	.32	-.17	.67	.35
UINT	-.14	1.00	.15	-.14	.06	.11
BIZK	.32	.15	1.00	-.06	.37	.22
POHT	-.17	-.14	-.06	1.00	-.59	-.53
HIPO	.67	.06	.37	-.59	1.00	.67
ESMT	.35	.11	.22	-.53	.67	1.00
	NPGB	UINT	BIZK	POHT	HIPO	ESMT

Hiru azpimatrize azpimarratu ditugu, 1. eta 2. aldagai ekonomikoak, 3. eta 4. populaziokoak, eta 5. eta 6. sozialak.

Bikote hauen erlazio mailak eta NPGB aldagaiak besteekin dituen erlazio mailak eta zeinuak asko esaten dute.

Koerlazio-matrizearen diagonalizazioaz egindako Osagai Nagusizko Analisian, ondoko $\lambda_1, ...\lambda_6$ autobalioak ditugu.

Autobalioak

EDITION DES VALEURS-PROPRES

SOMME DES VALEURS-PROPRES

HISTOGRAMME DES PREMIERS VALEURS-PROPRES

	VALEUR-PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	2.68648791	44.77	44.77	**********
2	1.14504230	19.08	63.86	**********
3	1.01897109	16.98	80.84	**********
4	.55648887	9.27	90.12	******
5	.43828008	7.30	97.42	*****
6	.15473048	2.58	100.00	**

Dakusagunez:

$$\sum_k \lambda_k = 6 = \text{tr } \mathbf{R}$$

hiru lehen ardatz faktorialetan proiektatutako bariantza % 80'84 da.
Hots: $\tau_3 = 80'84$

Aldagaien koordenatuak eta interpretaziorako laguntzak.

1 EDITION DES COORDONNEES ET DES CONTRIBUTIONS DES VARIABLES

NOMS	E.TYPE	* COORDONNES						* PROJECTION ANCIENS AXES UNITE (CARRE=CONTRIBUTION ABSOLUE)							* CORRELATION VARIABLE-FACTEUR (CARRE=CONTRIBUTION RELATIVE)						
		* F1	F2	F3	F4	F5	F6	* F1	F2	F3	F4	F5	F6	*	F1	F2	F3	F4	F5	F6	*
NPGB	639.238	* -.69	.53	-.13	.44	-.08	-.18	* -.42	.49	-.13	.59	-.12	-.45	*	-.69	.53	-.13	.44	-.08	-.18	*
UINT	142.013	* -.12	-.82	-.43	.35	.00	-.01	* -.08	-.77	-.42	.47	.00	-.03	*	-.12	-.82	-.43	.35	.00	-.01	*
BIZK	32960.844	* -.48	.11	-.76	-.42	-.07	-.04	* -.29	.10	-.75	-.56	-.11	-.10	*	-.48	.11	-.76	-.42	-.07	-.04	*
POHT	.676	* .68	.38	-.44	.20	.38	.13	* .41	.35	-.44	.27	.58	.32	*	.68	.38	-.44	.20	.38	.13	*
HIPO	16.410	* -.94	.08	.05	.10	-.07	.31	* -.57	.07	.05	.13	-.11	.80	*	-.94	.08	.05	.10	-.07	.31	*
ESMT	11,623	* -.80	-.17	.20	-.12	.52	-.08	* -.49	-.16	.20	-.17	.79	-.21	*	-.80	-.17	.20	-.12	.52	-.08	*

Aldagaien kasuan, $g_k(j)$ proiekzioak edo koordenatuak ditugu sei ardatzetarako, hau da $k \in \{1,..., 6\}$

Dakigunez, k. ardatz faktorialaren gain proiektatutako inertzia $\sum_i g_k^2(j) = \lambda_k$ da, kasu honetan $\mathbf{X}^T\mathbf{X} = \mathbf{R}$ diagonalizatzen denez, proiektatutako bariantza tipifikatua, hain zuzen.

Aldagai bakoitzari dagokion proiektatutako bariantzaren proportzioari aldagaiaren ekarpen absolutua deritzo. Hau da, $g_k^2(j)/\lambda_k$. aldagaiaren ekarpen absolutua da.

Dakusagunez, programak balio hauen erro karratuak ematen dizkigu edo $\mathbf{u}_k$ bektore unitarioen osagaiak.

Hots:

$$\mathbf{u}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{X}^T \mathbf{v}_k = \frac{\mathbf{g}_k}{\sqrt{\lambda_k}}$$

Azken sei zutabeetan aldagai eta faktoreen arteko koerlazioak ditugu, hauek $\mathbf{g}_k = \mathbf{X}^T \mathbf{v}_k$ bektoreen balioak direnez, aldagai normatuen eta $\mathbf{v}_k$ bektore unitarioaren arteko biderkadura eskalarrak, hain zuzen.

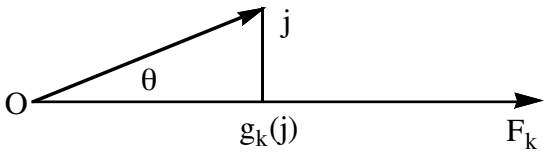
Balio hauek, $g_k^2(j)/d^2(j,0)$ balioen erro karratuak dira, analisi normatuan $d^2(j,0) = 1$ baita.

Ardatz faktorial bakoitzari dagokion aldagaiaren jatorriarekiko daukan distantzia karratuaren proportzioari, ardatzaren ekarpen erlatiboa deritzo.

Hau da, $g_k^2(j)/d^2(j,0)$ k. ardatzaren ekarpen erlatiboa da, faktoreak aldagaiari egindako ekarpen erlatiboa hain zuzen.

Dakigunez, $\sum_{k=1}^{6} g_k^2(j) = d^2(j,0)$ "Pitagoras-en Teorema Espazioan" kasu honetan $\mathbb{R}^{18}$ espazioan aplikatuz (aldagaien espazioa) lortutako sei ardatzetarako.

Ohartu behar dugu, $g_k(j) / d(j,0)$, aldagaiak ardatzarekin osatzen duen θ angeluaren kosinua dela, bien arteko koerlazioa, hain zuzen.



Hots:

$$\cos_k^2(\theta) = g_k^2(j)/d^2(j,0)$$

$$\sum_k \cos_k^2(\theta) = 1$$

Indibiduoen koordenatuak eta interpretaziorako laguntzak.

1 EDITION DES COORDONNEES ET DES CONTRIBUTIONS DES INDIVIDUS

NOMS	MASES	DISTO	*	COORDONNES						*	CONTRIBUTIONS ABSOLUES*100						*	CONTRIBUTIONS RELATIVES						*
			*	F1	F2	F3	F4	F5	F6	*	F1	F2	F3	F4	F5	F6	*	F1	F2	F3	F4	F5	F6	*
VEN	1.000	1.93	*	-.56	.94	-.23	.81	.03	.18	*	3.83	25.57	1.77	39.05	.05	6.90	*	.16	.46	.03	.34	.00	.02	*
ARG	1.000	1.63	*	-1.06	-.53	-.13	.43	-.11	-.04	*	13.93	8.26	.54	11.28	1.01	.37	*	.69	.17	.01	.12	.01	.00	*
MEX	1.000	.99	*	-.65	.36	-.57	-.15	.28	-.12	*	5.19	3.81	10.36	1.35	6.12	3.19	*	.42	.13	.33	.02	.08	.01	*
PAN	1.000	.41	*	-.11	.26	.30	.17	.13	-.44	*	.15	1.99	2.86	1.66	1.27	42.64	*	.03	.17	.21	.07	.04	.48	*
URU	1.000	1.89	*	-1.00	-.25	.78	-.10	-.45	.00	*	12.39	1.87	20.01	.65	15.20	.01	*	.53	.03	.32	.01	.11	.00	*
BRA	1.000	2.21	*	-.76	.06	-1.10	-.61	-.22	-.01	*	7.20	.11	39.64	1.95	3.72	.05	*	.26	.00	.55	.17	.02	.00	*
TXI	1.000	1.03	*	-.82	-.07	.52	-.16	.20	.16	*	8.30	.16	8.76	1.55	2.98	5.39	*	.65	.01	.26	.03	.04	.02	*
KOL	1.000	.20	*	-.18	.03	.18	-.18	-.31	.06	*	.40	.03	1.06	1.90	7.22	.84	*	.16	.00	.17	.16	.49	.02	*
CRI	1.000	.31	*	.51	.13	.08	.08	-.01	-.14	*	3.23	.52	.19	.40	.02	3.98	*	.84	.06	.02	.02	.00	.06	*
GUA	1.000	1.38	*	.94	.30	-.10	.10	-.62	-.04	*	11.03	2.67	.30	.63	29.11	.37	*	.64	.07	.01	.01	.28	.28	*
EKU	1.000	.26	*	.17	.06	.13	-.08	.45	-.07	*	.38	.12	.53	.35	15.25	.94	*	.12	.02	.06	.02	.77	.02	*
PER	1.000	.21	*	-.17	-.22	.13	-.15	.23	.19	*	.37	1.37	.52	1.43	4.15	8.10	*	.14	.23	.08	.11	.26	.18	*
PAR	1.000	.39	*	.52	-.02	.23	-.17	.03	-.20	*	3.36	.01	1.68	1.69	.07	8.75	*	.69	.00	.13	.07	.00	.10	*
NIK	1.000	.36	*	.51	.08	.06	-.02	.15	.25	*	3.24	.17	.11	.02	1.72	13.73	*	.73	.02	.01	.00	.06	.18	*
DOE	1.000	.24	*	.17	-.09	.31	-.25	.18	.05	*	.36	.26	3.21	3.84	2.56	.57	*	.12	.04	.42	.27	.14	.01	*
HON	1.000	1.01	*	.96	.18	-.06	-.07	.23	.06	*	11.32	.92	.12	.33	4.12	.74	*	.90	.03	.00	.01	.05	.00	*
ELS	1.000	1.03	*	.96	.12	-.02	-.09	-.26	.13	*	11.46	.43	.01	.45	5.08	3.42	*	.90	.01	.00	.01	.06	.02	*
BOL	1.000	2.53	*	.56	-1.33	-.50	.44	.07	-.01	*	3.85	.73	8.02	11.46	.35	.02	*	.12	.70	.10	.08	.00	.00	*

Indibiduoen kasuan, $d^2(i,0)$ jatorriarekiko daukan distantzia karratua daukagu, eta ondoren $\sqrt{m/n}\ f_k(i,0)$ koordenatu zuzenduak sei ardatzetarako.

Era berean, indibiduo bakoitzari dagokion proiektatutako bariantzaren proportzioari, indibiduoaren ekarpen absolutua deritzo.

Hau da, $f_k^2(i)/\lambda_k$ -i. indibiduoaren ekarpen absolutua da; kasu honetan, balio hauek ehunekotan ditugu.

Azken sei zutabeetan faktoreek indibiduoei egindako ekarpen erlatiboak ditugu.

Ekarpen erlatiboa, aldagaien kasurako definitu dugun kontzeptu berbera denez, hauxe da:

$$f_k^2(i)/d^2(i,0)$$

eta

$$\sum_{k=1}^{6} f_k^2(i) = d^2(i,0)$$

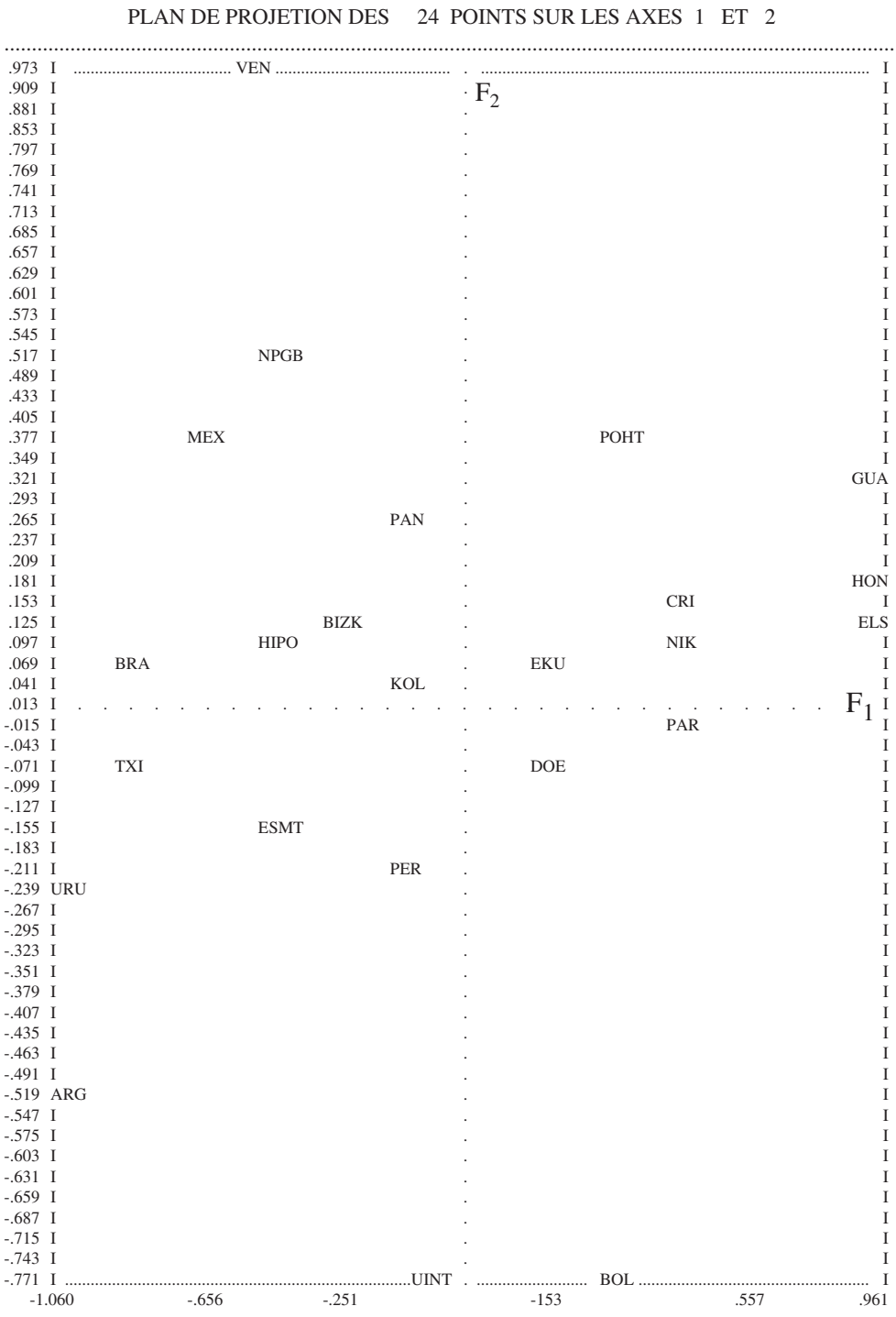
Indibiduoak ardatzarekin osatzen duen angeluaren kosinua $f_k(i)/d(i,0)$ da.

Hau da:

$$\cos_k^2(\theta) = f_k^2(i)/d^2(i,0)$$

$$\sum_k \cos_k^2(\theta) = 1$$

Baterako aurkezpen grafikoa: Plano Faktorial Nagusia



Grabitate-zentrutik gehien urruntzen diren aldagai eta indibiduoak dira koordenatu eta ekarpen absolutu eta erlatibo handienak dauzkatenak ardatz bakoitzerako, hau da, puntu horiek ardatz faktoriala aldagai berri bezala definitzera eramaten gaituzte.

Lehen ardatzaren kasuan HIPO (hiri-populazioaren ehunekoa) da aldagairik nabariena alde batean, bere koordenatua edota faktorearekin daukan koerlazioa -0'94 da eta honen arabera faktorearen bariantzari egindako ekarpen absolutua handiena da $(-0'57)^2$ (gogoratu behar dugu ardatzaren alde batean edo bestean egotea edonolakoa dela). ESMT (eskolako matrikula-tasa) eta NPGB (Nazio--Produktu Gordina) alde berean nabariak dira.

Aldagai hauen inguruan, planoaren ezker aldean, dauden indibiduorik nabarienak ARG, URU, TXI, eta atzetik MEX, BRA, dira. Azken bi hauek ekarpenetan dituzten ezberdintasunak aipagarriak dira; Brasilek faktoreari egindako ekarpen absolutua handiagoa dauka bere koordenatua handiagoa delako, baina faktoreak Brasili egindako ekarpen erlatiboa txikiagoa da Brasilek grabitate--zentrutik daukan distantzia handiagoa delako. Honek zera esan nahi du: Brasil beste ardatz batean hobeto aurkezten dela, kasu honetan, konkretuki, hirugarren ardatzarekiko dituen balioak adierazgarriagoak dira.

Ardatzaren beste aldean POHT (populazioaren hazkunde-tasa) da ekarpen absolutu eta erlatibo nabariaz daukagun aldagaia. Indibiduoen artean ELS, HON, GUA eta atzetik PAR, NIK, guzti hauek planoaren eskuin aldean ikus daitezkeelarik.

Lehen ardatza, bada, garapen-ardatza bezala defini daiteke, garapen-mota batena behintzat, eta honen arabera kontinentearen hegokonoa eta Ertamerika kontrajarrita daude.

Bigarren ardatzaren kasuan, UINT (urteroko inflazio-tasa) da aldagairik nabariena eta bere inguruan BOL eta ARG daude, ardatzean proiektatzen den bariantzaren erdia baino gehiago Boliviari dagokiolarik. Puntu hauek planoaren beheko aldean ikus daitezke.

Ardatzaren beste aldean, hau da, planoaren goiko aldean, NPGB eta POHT daude, biak lehen ardatzean kontrajarrita eta koerlazio hobeagorekin genituenak eta hauen inguruan VEN, ardatzean proiektatutako bariantzaren laurdena berari dagokiona, eta MEX, azken honek ekarpen absolutu eta erlatiboak txikiak dituelarik.

Bigarren ardatza edo beste edozein beti hondar-ardatza da aurrenekoekiko; zentzu honetan bigarrenak lehenengoak azaltzen ez dituen datu-matrizearen sakabanatzearen alderdiak azaltzen ditu.

Adibide honen kasuan bigarren ardatzak, gehienbat, kontinente honetan hain berezia den inflazio-tasaren arabera dauden desberdintasunak azaltzen ditu. Halere, Nazio-Produktu Gordina eta populazioaren hazkunde–tasa aldagaien konbinazio lineala den bigarren faktorean garrantziaz sartu dira.

Hirugarren ardatzean populazioko aldagaiak dira garrantzitsuenak, nabariena BIZK (biztanle kopurua) izanik. Faktore honen arabera populazio handia edo gehikorra duten estatuak (BRA, MEX) eta populazio ez hain handia eta gehikorra dutenak (URU, TXI) kontrajartzen dira. Azkenik esan behar dugu, interpretatu nahi izan dugun hirugarren ardatza beste biekiko hondar-ardatza dela. Datu-matrizearen desberdintasun globalik garrantzitsuenak lehen ardatzean proiektatzen ziren, sakabanatzearen % 44'77 jasoz eta bigarren garrantzitsuenak eta ez hain globalak bigarrenean, sakabanatzearen % 19'08 jasoz, hain zuzen ere.

IX.4. KORRESPONDENTZI ANALISI FAKTORIALA

Datu-Analisiaren eskola frantsesaren aita bezala ezaguna den J.P. Banzécri-k Korrespondentzi Analisi Faktoriala (KAF) garatu zuen 60.eko hamarkadaren lehen urteetan.

Estatistika inferentzialetik at lehen ezer gutxi ikasiak izan ziren kontingentzi tauletan arrakasta handiz aplikatzen da orduez gero KAF-ren metodologia[1] eta honek Datu-Analisiaren eskolaren bultzada suposatu zuen.

Teknika honi buruz egin daitezkeen aurkezpen ezberdinekin loturik daude KAF-ren helburuak.

Halere, helburu orokor bat kontsidera daiteke: populazio baten gain ohartutako bi aldagaien arteko menpekotasun-erlazioen azterketa egitea, hain zuzen.

KAF taula-mota ezberdinetara aplikagarria da, kontingentzi tauletara, alegia, eta zentzu zabalean kontsideraturik kontingentzi taulak bezala har daitezkeen tauletara.

(1) K. FERNANDEZ AGIRRE. *Análisis multivariante de tablas de gran inercia. Aplicación al corpus con terminologia económica en euskara*. Doktoradutza-tesia (1.988)

Kontingentzi taula

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & \dots & j & \dots & n \end{array} \\ \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1j} & \dots & k_{1n} \\ k_{21} & k_{22} & \dots & k_{2j} & \dots & k_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ k_{i1} & k_{i2} & \dots & k_{ij} & \dots & k_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ k_{m1} & k_{m2} & \dots & k_{mj} & \dots & k_{mn} \end{bmatrix} \end{array} = \left[k_{ij} \right]$$

i. errenkadaren eta j. zutabearen ebakiduran daukagun k_{ij} balioa, kontaketa baten emaitza izango da, normalki, (i,j) bikotearen maiztasun absolutua.

Halere, zentzu zabalean, kontingentzi taula definituz gero, errenkadaren baturak, $k_{i.}$ balioak hain zuzen, eta zutabeen baturak, $k_{.j}$ balioak, zentzu argia daukaten balioak izan daitezke nahiz eta berez taula horiek neurri taulak izan. Adibidez, enpresa batzuren salmenta kopuruak, esportatutako balioak... izan daitezke balio horiek.

Askotan, kontingentzi taulei menpekotasun taulak edo taula gurutzatuak deritze eta batzutan maiztasun-taulak ere. (Ikus III. ikasgaia).

Taula hauetan, errenkadak nahiz zutabeak populazio beraren partizioak dira, atribututan kategoriatan edo balio ordinaletan egindako partizioak alegia, horrela errenkadek nahiz zutabeek betebehar berdinak dauzkate.

Kontingentzi taulatik errenkada-soslaiak eta zutabe-soslaiak eratortzen dira errenkada-puntuen eta zutabe-puntuen arteko distantziek zentzua eduki dezaten.

Honela, errenkada-soslaien taulan, $k_{i.}$ errenkadaren totalarekiko proportzioak eta zutabe-soslaien taulan, $k_{.j}$ zutabearen totalarekiko proportzioak ditugu.

Hots:

$k = \sum_i \sum_j k_{ij}$ kontingentzi taularen totala

$f_{ij} = k_{ij} / k$ maiztasun erlatiboak

$$\left.\begin{array}{l} f_{i.} = \sum_i f_{ij} \\ f_{.j} = \sum_j f_{ij} \end{array}\right\}$$ bazter-maiztasunak

Errenkada-soslaien taula

	1	2	...	j	...	n	Totala
1	$\frac{f_{11}}{f_{1.}}$	$\frac{f_{12}}{f_{1.}}$	...	$\frac{f_{1j}}{f_{1.}}$	...	$\frac{f_{1n}}{f_{1.}}$	1
2	$\frac{f_{21}}{f_{2.}}$	$\frac{f_{22}}{f_{2.}}$	...	$\frac{f_{2j}}{f_{2.}}$	...	$\frac{f_{2n}}{f_{2.}}$	1
.....							
i	$\frac{f_{ij}}{f_{1.}}$	$\frac{f_{i2}}{f_{1.}}$	...	$\frac{f_{ij}}{f_{i.}}$	...	$\frac{f_{in}}{f_{i.}}$	1
.....							
m	$\frac{f_{m1}}{f_{m.}}$	$\frac{f_{m2}}{f_{m.}}$	...	$\frac{f_{mj}}{f_{m.}}$	...	$\frac{f_{mn}}{f_{m.}}$	1
Totala	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.n}$	

Zutabe-soslaien taula

	1	2	...	j	...	n	Totala
1	$\frac{f_{11}}{f_{.1}}$	$\frac{f_{12}}{f_{.2}}$	...	$\frac{f_{1j}}{f_{.j}}$	...	$\frac{f_{1n}}{f_{.n}}$	$f_{1.}$
2	$\frac{f_{21}}{f_{.1}}$	$\frac{f_{22}}{f_{.2}}$	...	$\frac{f_{2j}}{f_{.j}}$	...	$\frac{f_{2n}}{f_{.n}}$	$f_{2.}$
.....							
i	$\frac{f_{i1}}{f_{.1}}$	$\frac{f_{i2}}{f_{.2}}$	...	$\frac{f_{ij}}{f_{.j}}$	...	$\frac{f_{in}}{f_{.n}}$	$f_{i.}$
.....							
m	$\frac{f_{m1}}{f_{.1}}$	$\frac{f_{m2}}{f_{.2}}$	...	$\frac{f_{mj}}{f_{.j}}$	...	$\frac{f_{mn}}{f_{.m}}$	$f_{m.}$
Totala	1	1	...	1	...	1	

IX.4.1. Hodeiak, masak eta distantziak

Osagai Nagusizko Analisian bezala, bi espazio desberdinetan koka gaitezke.

a) $\mathbb{R}^n$ espazio bektorialean m errenkada-soslaien puntuak ditugu. Honela, $f_{i.}$ masaz hornitutako i puntuaren koordenatuak ondoko multzokoak dira:

$$\left\{\left(f_{ij}/f_{i.}\right) \;;\; j = 1,2,\ldots.n \right\}$$

Konkretuki, errenkada-soslaien hodeiaren m puntuak $\mathbb{R}^{n-1}$ azpiespazio batean kokaturik daude, puntu bakoitzaren koordenatuen batura 1 baita.

Hau da:

$$\sum_j \left(f_{ij}/f_{i.}\right) = 1 \qquad \forall i = 1,2,\ldots.m$$

Honela, puntuen hurbiltasunak soslaien hurbiltasunak bezala interpreta daitezke.

b) $\mathbb{R}^m$ espazio bektorialean n zutabe-soslaien puntuak ditugu. Honela, $f_{.j}$ masaz hornitutako j puntuaren koordenatuak ondoko multzokoak dira:

$$\left\{\left(f_{ij}/f_{.j}\right) \;;\; i = 1,2,\ldots.m \right\}$$

Konkretuki, zutabe-soslaien hodeiaren n puntuak $\mathbb{R}^{m-1}$ azpiespazio batean kokaturik daude, puntu bakoitzaren koordenatuen batura 1 baita.

Hau da:

$$\sum_i \left(f_{ij}/f_{i.}\right) = 1 \qquad \forall j = 1,2,\ldots.n$$

Doikuntza egin aurretik puntu bakoitzaren masa kontutan hartzea guztiz zilegia da; horrela, masa txikiak dituzten puntuak ez dira gehiegi nagusitzen eta banaketa bere baitan errespetatzen da.

$\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioetan definitzen diren distantziak hauexek dira:

$$d^2(i,i') = \sum_{j=1}^{n} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

$$d^2(j,j') = \sum_{i=1}^{m} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

Horrela definituriko distantziari χ^2 (ji karratua) deritzo.

Distantzia ponderatu honek daukan abantaila, guztiz berezkoa izateaz gain, **banaketa-baliokidetasun** izeneko propietatea betetzea da.

Propietate honen arabera i_1, i_2 errenkada-puntuak soslai berbera badira (proportzio berdinak zutabe guztietarako) berorien masen batura daukan i_o errenkada-puntuaz ordezka daitezke eta $\mathbb{R}^n$ nahiz $\mathbb{R}^m$ espazioetan distantziak ez dira aldatzen. Berdin esan daiteke j_1, j_2 soslai berbereko bi zutabe-puntuentzako.

Propietate honek analisiaren emaitzak oso egonkor bihurtzen ditu; honela, datu-taula handi batean berdin antzeko soslaia daukaten lerroak emaitzak mantenduz batu daitezke eta honek nomenklatura baten arbitraltasunaren aurrean emaitzen garrantzia suposatzen du.

Soslaien artean definitutako χ^2 distantzia ohizko distantzia euklidear bihur daiteke, ponderazioak parentesien barruan sartuz.

Hau da:

$$d^2(i,i') = \sum_{j=1}^{n} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j'}}{f_{i'.}\sqrt{f_{.j}}} \right)^2$$

Kasu honetan, ardatzen eskala aldatuz i puntuaren koordenatuak $f_{ij}/f_{i.}\sqrt{f_{.j}}$ balioak dira.

Era berean:

$$d^2(j,j') = \sum_{i=1}^{m} \left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \frac{f_{ij'}}{f_{.j'}\sqrt{f_{i.}}} \right)^2$$

Kasu honetan, ardatzen eskala aldatuz j puntuaren koordenatuak $f_{ij}/f_{.j}\sqrt{f_{i.}}$ balioak dira.

Eskala aldatu ondoren, errenkada-soslaien grabitate-zentrua zutabeen bazter--banaketaren erro karratua da.

i puntu bakoitzaren masa edo pisua $f_{i.}$ denez:

$$g_j = \sum_{i=1}^{n} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} f_{i.} = \sqrt{f_{.j}}$$

Eta zutabe-soslaien grabitate-zentrua errenkaden bazter-banaketaren erro karratua da. j puntu bakoitzaren masa edo pisua $f_{.j}$ denez:

$$h_i = \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}\sqrt{f_{.j}}} f_{.j} = \sqrt{f_{i.}}$$

Azkenik:

$$\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \text{ eta } \frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \sqrt{f_{i.}}$$

bi hodeien koordenatu zentratuak, hurrenez hurren, dira.

Dakusagunez, bada, puntu-hodeiaren grabitate-zentrua jatorrira eraman dugu eta ardatzen eskala-aldaketa baten bidez χ^2 distantzia ohizko distantzia euklidearra bezala idatzi dugu, honela, korrespondentzi analisi faktoriala analisi orokorraren barrutian aurkezten da.

IX.4.2. $\mathbb{R}^n$ espazioan egindako analisia

$\mathbb{R}^n$ espazioan diagonalizatzen den kobariantza matrizearen gai orokorra i puntuaren masa $f_{i.}$ dela kontutan hartuz, hauxe da:

$$S_{jj'} = \sum_{i=1}^{m} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \right) \left(\frac{f_{ij'}}{f_{i.}\sqrt{f_{.j'}}} - \sqrt{f_{.j'}} \right) f_{i.}$$

j, j' puntuen balio zentratuen arteko biderkadura da i guztietarako eta i puntuen masaz ponderaturik.

Azkenik, biderkatzaile bakoitza izendatzaile komunean ipiniz eta $\sqrt{f_{i.}}$ biderkatuz.

$$S_{jj'} = \sum_{i=1}^{m} \left(\frac{f_{ij} - f_{i.}\, f_{.j}}{\sqrt{f_{i.}\, f_{.j}}} \right) \left(\frac{f_{ij'} - f_{i.}\, f_{.j'}}{\sqrt{f_{i.}\, f_{.j'}}} \right)$$

Horrela diagonalizatzen den **L**(J) kobariantza matrizea **X**(m,n) matrizearen funtzioan idatz daiteke:

$$\underset{(n,n)}{\mathbf{L}(J)} = \underset{(n,m)}{\mathbf{X}^T}\ \underset{(m,n)}{\mathbf{X}}$$

non, **X** matrizearen gai orokorra $\dfrac{f_{ij} - f_{i.}\, f_{.j}}{\sqrt{f_{i.}\, f_{.j}}}$ da.

Hots:

$$\mathbf{X} = \left[\frac{f_{ij} - f_{i.}\, f_{.j}}{\sqrt{f_{i.}\, f_{.j}}} \right]$$

Froga daiteke datuak zentratuz edo zentratu gabe erabiltzeak emaitza berdinetara eramaten duela.

Hau da, datuak zentratu gabe hartzen baditugu

$$\mathbf{X}^* = \left[\frac{f_{ij}}{\sqrt{f_{i.}\, f_{.j}}} \right]$$

eta kobariantza matrizea $\mathbf{L}^*(J) = \mathbf{X}^{*T}\mathbf{X}^*$ izango da.

$\mathbf{u}_n$ bektorea, non j. osagaia $\sqrt{f_{.j}}$ baita, $\mathbf{L}(J)$ kobariantza matrizearen 0 autobalioari dagokion autobektorea da; halaber, $\mathbf{u}_n$ (grabitate-zentrua) $\mathbf{L}^*(J)$ kobariantza matrizearen 1 autobalioari dagokion autobektorea da.

$\mathbf{u}_n$ ez den edozein $\mathbf{u}_k$ autobektore, $\mathbf{L}(J)$ edota $\mathbf{L}^*(J)$ matrizeen autobalio berberari dagokion autobektorea da.

k. ardatzaren gain m errenkada-puntuek dauzkaten proiekzioak, datu zentratuaz nahiz zentratu gabeaz kalkula daitezke,

$$\underset{(m,1)}{\hat{\psi}_k} = \underset{(m,n)(n,1)}{\mathbf{X}\ \mathbf{u}_k} = \underset{(m,n)(n,1)}{\mathbf{X}^*\mathbf{u}_k}$$

$\mathbf{X}\,\mathbf{u}_n$ eta $\mathbf{X}^*\,\mathbf{u}_n$ bektoreen abzisak berberak baitira.

Dakusagun i puntuaren koordenatua bektore batean nahiz bestean berbera dela.

$$\hat{\psi}_{ki} = \sum_{j=1}^{n}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}}\right) u_{kj} =$$

$$= \sum_{j=1}^{n}\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}\,u_{kj} \qquad (8)$$

Parentesia kentzean bigarren batugaia $\sum_{j=1}^{n}\sqrt{f_{.j}}\ u_{kj} = 0$ da, $\mathbf{u}_n$ eta $\mathbf{u}_k$ autobektoreak ortogonalak baitira.

IX.4.3. $\mathbb{R}^m$ espazioan egindako analisia

$\mathbb{R}^m$ espazioan diagonalizatzen den kobariantza matrizearen gai orokorra j puntuaren masa $f_{.j}$ dela kontutan hartuz, hauxe da:

$$S_{ii'} = \sum_{j=1}^{n}\left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \sqrt{f_{i.}}\right)\left(\frac{f_{i'j}}{f_{.j}\sqrt{f_{i'.}}} - \sqrt{f_{i'.}}\right) f_{.j}$$

Eta biderkatzaile bakoitza izendatzaile komunean ipiniz eta $\sqrt{f_{.j}}$-z biderkatuz

$$S_{ii'} = \sum_{j=1}^{n} \left(\frac{f_{ij} - f_{i.} f_{.j}}{\sqrt{f_{i.} f_{.j}}} \right) \left(\frac{f_{i'j} - f_{i'.} f_{.j}}{\sqrt{f_{i'.} f_{.j}}} \right)$$

Hau da:

$$\underset{(n,n)}{\mathbf{L}(I)} = \underset{(m,n)}{\mathbf{X}} \quad \underset{(n,m)}{\mathbf{X}^T}$$

Dakusagunez korrespondentzian jartzen diren errenkaden eta zutabeen multzoek betebehar berdinak dauzkate eta espazio batetik bestera pasatzeko i,j azpiindizeen betebeharrak aldatzea aski da.

$\mathbf{v}_m$ bektorea, non i osagaia $\sqrt{f_{i.}}$ den, $\mathbf{L}(I)$ kobariantza matrizearen 0 autobalioari dagokion autobektorea da; halaber, $\mathbf{v}_m$ (grabitate-zentrua), $\mathbf{L}^*(I) = \mathbf{X}^*\mathbf{X}^{*T}$ kobariantza matrizearen 1 autobalioari dagokion autobektorea da.

$\mathbf{v}_m$ ez den edozein $\mathbf{v}_k$ autobektore, $\mathbf{L}(I)$ edota $\mathbf{L}^*(I)$ matrizeen autobalio berberari dagokion autobektorea da.

Azkenik, k. ardatzaren gain n zutabe-puntuek dauzkaten proiekzioak:

$$\hat{\varphi}_k = \mathbf{X}^T \mathbf{v}_k = \mathbf{X}^{*T} \mathbf{v}_k$$

Eta j puntuaren koordenatua bere adierazpen errazenean:

$$\hat{\varphi}_{kj} = \sum_{i=1}^{m} \frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} v_{ki} \qquad (9)$$

IX.4.4. $\mathbb{R}^n$ eta $\mathbb{R}^m$ espazioen arteko erlazioa

Analisi orokorrean ikusi genuenez $\mathbf{L}^*(J) = \mathbf{X}^{*T}\mathbf{X}^*$ eta $\mathbf{L}^*(I) = \mathbf{X}^*\mathbf{X}^{*T}$ matrize simetrikoek desberdin zero diren λ_k autobalio berberak dituzte eta λ_k autobalioari dagozkion $\mathbf{L}^*(J)$-ren $\mathbf{u}_k$ eta $\mathbf{L}^v(I)$-ren $\mathbf{v}_k$ autobektoreen arteko erlazioak hauexek dira:

$$\mathbf{v}_k = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}^* \mathbf{u}_k \qquad (10)$$

$$\mathbf{u}_k = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}^{*T} \mathbf{v}_k \qquad (11)$$

$\mathbf{v}_k$ bektorearen i. errenkada (10) kontutan hartuz, honela idatz daiteke:

$$v_{ki} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} \frac{f_{ij}}{\sqrt{f_{i.}\, f_{.j}}}\, u_{kj}$$

Eta (8) erlazioan u_{kj} ordezkatuz:

$$\hat{\psi}_{ki} = \sqrt{\lambda_k}\, \frac{v_{ki}}{\sqrt{f_{i.}}} \qquad (12)$$

Era berean, $\mathbf{u}_k$ bektorearen j. errenkada (11) kontutan hartuz honela idatz daiteke:

$$u_{kj} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m} \left(\frac{f_{ij}}{\sqrt{f_{i.}\, f_{.j}}} \right) v_{ki}$$

Eta (9) erlazioan v_{ki} ordezkatuz:

$$\hat{\varphi}_{kj} = \sqrt{\lambda_k}\, \frac{u_{kj}}{\sqrt{f_{.j}}} \qquad (13)$$

Azkenik, (9) erlazioan (12) erlaziotik ateratako v_{ki} ordezkaturik:

$$\hat{\varphi}_{kj} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m} \frac{f_{ij}}{f_{.j}}\, \hat{\psi}_{kj} \qquad (14)$$

eta (8) erlazioan (13) erlaziotik ateratako u_{kj} ordezkaturik:

$$\hat{\psi}_{ki} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} \frac{f_{ij}}{f_{i.}}\, \hat{\varphi}_{kj} \qquad (15)$$

(14), (15) erlazioetan errenkada-puntuek eta zutabe-puntuek k. ardatzaren gain dituzten koordenatuen arteko erlazioak ditugu.

Dakusagunez (f_{ij} / $f_{i.}$) gai orokorraz daukan matrizearen bidez, hau da, errenkada-soslaien matrizearen bidez, eta j zutabe-puntuen koordenatuen bidez ((15) erlazioa), i errenkada-puntuen koordenatuak lortzen dira.

Honela, i errenkada-puntuaren koordenatua k. ardatzaren gain $\left(\hat{\psi}_{ki}\right)$, $1/\sqrt{\lambda}$ balioa salbu, j zutabe-puntuen koordenatuen barizentrua edo grabitate-zentrua da, pisuak i errenkada-soslaiaren balioak direlarik.

Era berean, j zutabe-puntuaren koordenatua k. ardatzaren gain $\left(\hat{\varphi}_{kj}\right)$, $1/\sqrt{\lambda}$ balioa salbu, i errenkada-puntuen koordenatuen barizentrua da, pisuak j zutabe--soslaiaren balioak direlarik ((14) erlazioa).

Erlazio barizentrikoak bezala ezagutzen diren (14), (15) erlazioak ez dira analisi orokorraren (5), (6) erlazioen kasu bereziak, KAF-ean trantsizio-erlazioaren matrizeak ez baitira bata bestearen iraule.

Erlazio hauetan, inplizituki, edozein k-rako $\lambda_k \leq 1$ dela daukagu $\left(1/\sqrt{\lambda_k}\right) < 1$ ezin baita izan.

IX.4.5. Maiztasun-taularen berreraketa

Gai honen 2.4. atalean, analisi orokorrerako ikusi genuenez, $\mathbf{X}^* = \left[f_{ij}/\sqrt{f_{i.}\ f_{.j}}\right]$ matrizearen berreraketa hauxe da:

$$\mathbf{X}^* = \sum_{k=1}^{n} \sqrt{\lambda_k}\ \mathbf{v}_k\ \mathbf{u}_k^T$$

(12), (13) erlazioetatik ateratako $\mathbf{v}_k$, $\mathbf{u}_k$ balioak ordezkatuz gero:

$$\frac{f_{ij}}{\sqrt{f_{i.}\ f_{j}}} = \sum_{k=1}^{n}\left[\frac{1}{\sqrt{\lambda_k}}\ \psi_{ki}\ \varphi_{kj}\right]\sqrt{f_{i.}\ f_{.j}}$$

Eta dakigunez $\lambda_1 = 1$ autobalioari $\mathbf{u_{1j}} = \sqrt{f_{.j}}$, $\mathbf{v_{1i}} = \sqrt{f_{i.}}$ autobektoreak dagozkio, orduan, (12), (13) erlazioetan dakusagunez $\hat{\varphi}_{1j} = 1$ eta $\hat{\psi}_{1i} = 1$

$$f_{ij} = f_{i.}\ f_{.j}\left[1 + \sum_{k=2}^{n}\frac{1}{\sqrt{\lambda_k}}\ \psi_{ki}\ \varphi_{ki}\right]$$

Berreratze-formula honetan, $\hat{\psi}_{ki}$, $\hat{\varphi}_{ki}$ eta λ_k hirurek zerorantz jotzen dute.

Askotan, berreratze-formula, faktore normatuen bidez idazten da ere.

ψ_{ki} eta φ_{ki} faktore normatuak:

$$\psi_{ki} = \frac{1}{\sqrt{\lambda_k}} \hat{\psi}_{ki} \quad \text{eta} \quad \varphi_{ki} = \frac{1}{\sqrt{\lambda_k}} \hat{\varphi}_{ki}$$

Orduan berreratze-formula:

$$f_{ij} = f_{i.}\, f_{.j} \left[1 + \sum_{k>1} \sqrt{\lambda_k}\, \psi_{ki}\, \varphi_{kj} \right]$$

IX.4.6. Interpretaziorako laguntzak

Osagai Nagusizko Analisiaren adibidean ikusi genuen bezala, ardatzen interpretaziorako, bi koefiziente-segidak kalkulatzen dira korrespondentzian ipinitako bi multzoen elementu bakoitzerako.

1) *Ekarpen absolutuak*

Dakigunez, elementu bakoitzari dagokion proiektatutako bariantzaren proportzioari elementuaren ekarpen absolutua deritzo.

k. ardatzaren gain m errenkada-puntuen bariantza proiektatua, i errenkada-puntu bakoitzaren masa $f_{i.}$ dela jakinik, hauxe da:

$$\sum_{i=1}^{m} f_{i.}\, \hat{\psi}_{ki}^2$$

Eta $\hat{\psi}_{ki}$ balioa ordezkatuz:

$$\sum_{i=1}^{m} f_{i.}\, \hat{\psi}_{ki}^2 = \lambda_k \sum_{i=1}^{m} v_{ki}^2 = \lambda_k \quad (\mathbf{v}_k \text{ unitate bektorea baita})$$

Era berean j zutabe-puntuen bariantza proiektatua.

$$\sum_{j=1}^{n} f_{.j}\ \hat{\varphi}_{kj}^2 = \lambda_k \sum_{j=1}^{n} u_{kj} = \lambda_k$$

Honela, k. ardatzaren gain proiektatutako bariantzan i elementuari dagokion proportzioa:

$$Ca_k\ (i) = f_{i.}\ \hat{\psi}_{ki}^2 / \lambda_k$$

Ca_k (i) balioa, i elementuak k. ardatzean daukan ekarpen absolutua izanik.

Ohar daiteke:

$$\sum_{i=1}^{m} Ca_k(i) = 1 \qquad \forall_k \text{ dela}$$

Analogikoki, j elementuak k. ardatzean daukan ekarpen absolutua:

$$Ca_k(j) = \frac{f_{.j}\ \hat{\varphi}_{kj}^2}{\lambda_k}$$

2) *Ekarpen erlatiboak*

Dakigunez, ardatz faktorial bakoitzari dagokion elementuaren jatorriarekiko daukan distantzia karratuaren proportzioari, ardatzaren ekarpen erlatiboa deritzo.

Ardatz faktorialek, bi espaziotan, oinarri ortonormalak osatzen dituztenez, bi espazioko G, H, grabitate-zentruekiko elementuek dituzten distantzia karratuak ardatzetan dituzten koordenatuen karratuetan deskonposatzen dira batukorki.

Honela, $\mathbb{R}^n$ espazioko i puntu bakoitzerako:

$$d_n^2(i,G) = \sum_{j=1}^{n} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \right)^2$$

Halaber, $\mathbb{R}^m$ espazioko j puntu bakoitzerako:

$$d_m^2(j,H) = \sum_{i=1}^{m} \left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \sqrt{f_{i.}} \right)^2$$

Puntu baten soslaiak batezbesteko soslaiarekin kointziditzen badu, distantzia zero izango dela ohartu behar dugu.

k. ardatzaren ekarpen erlatiboa i puntuarekiko:

$$Cr_k\ (i) = \frac{\hat{\psi}^2_{ki}}{d^2_n\,(i,G)}$$

eta i puntuak k. ardatzarekin osatzen duen angeluaren cosinu karratua denez, koerlazio-koefiziente karratua bezala interpreta daiteke.

Ohar daiteke

$$\sum_k Cr_k\ (i) = 1 \qquad \forall_i \text{ dela}$$

Analogikoki, k. ardatzak j elementuaren posizioarekiko daukan ekarpen erlatiboa:

$$Cr_k\ (j) = \frac{\hat{\varphi}^2_{kj}}{d^2_m\,(j,H)}$$

Eta, era berean:

$$\sum_k Cr_k\ (j) = 1 \qquad \forall\, j$$

IX.4.7. Korrespondentzia Anitzeko Analisi Faktoriala

Korrespondentzi Analisi Faktorialaren orokorpena bezala Korrespondentzia Anitzeko Analisi Faktoriala sortzen da.

Bi aldagai baino gehiago, kualitatiboak nahiz kuantitatiboak izan, azken kasu honetan bada kategoriatan mailakatuta, ditugunean, Korrespondentzia Anitzeko Analisi Faktorialaren metodologia analisi ikuskatzailea egiteko guztiz baliagarria da.

Teknika honen aplikazio-eremu handi bat galdera itxien inkestetakoa da; halaber, edozein inkesta motatan lortutako erantzunak aztertzeko, oso egokia da.

Korrespondentzia Anitzeko Analisiaren eragiketaren teknika, deskribapen logikozko taula batean Korrespondentzi Analisi Faktoriala aplikatzea besterik ez da.

Deskribapen logikozko taula

Taula hauen elementuak 0 eta 1 dira. k_{ij} taularen gai orokorra izanik, i elementuak j propietatea duenean $k_{ij} = 1$ izango da, aldiz, i elementuak j propietaterik ez duenean $k_{ij} = 0$ izango da.

Deskribapen logikozko taula batek ondoko baldintza betetzen baldin badu aukera osoko taularen erara dagoela esango dugu: J zutabeen (edo modalitateen) multzoa Jq azpimultzoen (edo galderen) Q segida da non:

$$\forall\, i \in I,\ \forall\, J_q \in Q, \exists\, j \in J_q \mid k_{ij} = 1 \wedge (j' \in J_q\,.\, j' \neq j) \Rightarrow k_{ij'} = 0$$

Honek zera esan nahi du: I indibiduoen multzoko i elementu bakoitzak Jq klasean modalitate bat eta bakarrik bat edukiko duela.

Deskribapen logikozko taula batek aukera osoko taularen erara honelako itxura dauka:

	J_1	....	J_2	....	J_q	....	J_Q	Totala
1	010	....	1000	....	0010	....	010	Q
.	.....							
.	.....							
Z = i	001	....	0100	....	0100	....	100	Q
(m,J) .	.....							
.	.....							
m	100	....	0001	....	0100	....	010	Q

Dakusagunez i indibiduoak Jq azpimultzoan bigarren modalitatea aukeratu du; inkesta baten kasua balitz, q. galderarentzako bigarren erantzuna aukeratu duela esango genuke.

Korrespondentzi Anitzeko Analisia eta Burt-en taula batetan Korrespondentzi Analisi Faktoriala aplikatzea baliokideak dira.

Burt-en taula

m indibiduoen gain Q aldagai (edo galdera) ohartu baditugu, Burt-en taula (J_q, $J_{q'}$) binaka gurutzatutako taula guztiez osaturik dago.

B =

	J_1	J_2	…	J_q	…	J_Q
J_1	0 / 0	J_2		J_q		J_Q
J_2	J_{21}	0 / 0		J_{2q}		J_{2Q}
⋮						
J_q	J_{q1}	J_{q2}		0 / 0		J_{qQ}
⋮						
J_Q	J_{Q1}	J_{Q2}		J_{Qq}		0 / 0

Diagonalean ditugun azpitaulak bloke diagonalak dira; honelako (J_q, J_q) blokeek, diagonalean, aukeratutako modalitateen kopuruak dituzte eta beste elementuak zero dira, bi modalitate (edo erantzun) desberdin batera ezin baitira aukeratu.

(J_q, $J_{q'}$) blokeak, benetako kontingentzia taulak dira, bi galderen erantzunak gurutzatu baitira.

B deitutako Burt-en taula, matrizialki, **Z** taularen bidez honela lortuko da:

$$\mathbf{B} = \mathbf{Z}^T \mathbf{Z}$$

Korrespondentzi Analisia aplikatu ondoren lortutako faktoreak interpretatzerako, KAF-ean lortutako erlazioak eta interpretaziorako laguntzak (ekarpen absolutuak eta erlatiboak) baliagarriak dira.

IX.5. ELEMENTU GEHIGARRIAK

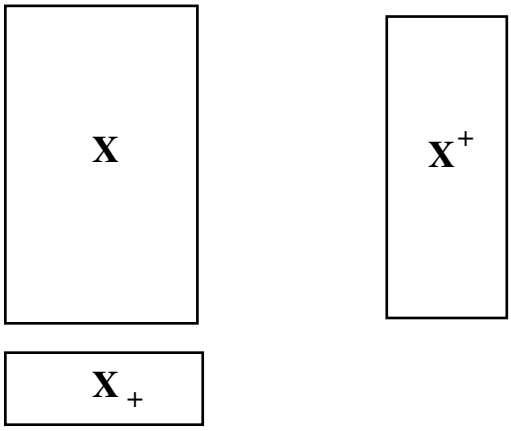
Korrespondentzi Analisi Faktorialean (bikoitza nahiz anizkoitza) eta Osagai Nagusizko Analisian *elementu gehigarrien teknikak* asko laguntzen du analisiaren interpretazioan.

Laginetik at, n aldagaien gain ditugun indibiduo batzuen balioak ezaguturik, analisiaren lekuko bezala, hodeian, talde hori kokatzea komenigarria gerta daiteke.

Batzutan, ezaugarri interesgarri bat daukaten indibiduo-multzoak izanik, berorien grabitate-zentruak indibiduo gehigarri bezala hartuko dira. Azkenik, jatorrizko laginerako beste aldagai batzuen balioak baztertuak izanik, datu--multzoan homogenoak ez direlako gehigarri bezala aurkeztea komenigarria izatea gerta daiteke.

Arazoa, bada, hauxe da: Errenkada-puntu edota zutabe-puntu horien koordenatu faktorialak lortzea, honela ardatz eta plano faktorialetan aurkeztu ahal dira.

Eskematikoki,



X^+ nahiz X_+ tauletan ditugun elementuen koordenatu faktorialak, elementu aktiboenak bezala oso erraz kalkulatuko dira eta eskola frantsesaren datu-analisiaren esparruan egiten diren estatistikaren programa-paketetan ahalbide bezala aurki daitezke.

Elementu hauek nahiz eta analisian, aktiboki, parterik hartu ez, ardatz eta planotan proiektatzean, batzuetan, kokapen pribilegiatua edukiz asko lagundu ahal dute emaitzen azalpenetan.

IX.6. SAILKAPEN METODOAK

Gaiaren hasieran esaten genuenez estatistika deskribatzaile anizkoitzean bi analisi-mota egin daitezke eta biak datu-hodei berberan aplikagarriak dira.

Orain arte ikusi ditugun analisi faktorialen bidez hodeiaren direkzio luzatuenak bilatzen dira eta beraien gain bildutako sakabanatzerik adierazgarriena azaltzen da.

Sailkapen metodoen bidez, ordea, hodeiaren elementu (indibiduo) hurbilenak sailetan kokatzen saiatuko gara. Lortutako sailak beraien elementutan homogenoak izango dira eta sailen arteko desberdintasunak ahal den hoberen finkatuko dira.

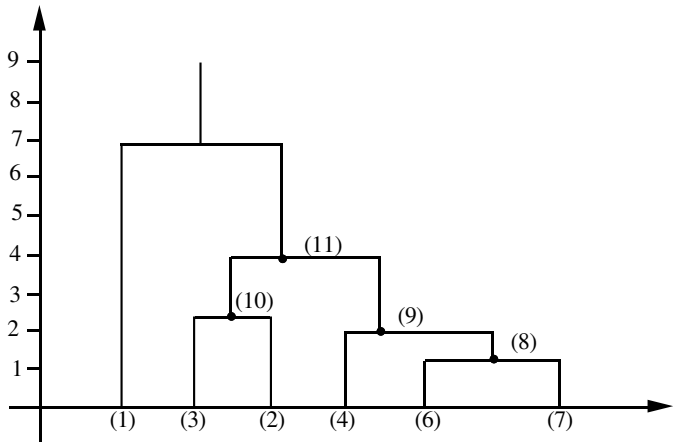
Normalki, sailkapen metodoak analisi faktorialaren emaitzak osotzeko eta aberasteko erabiltzen dira; kasu hauetan indibiduoek dituzten lehen koordenatu faktorialetatik abiatuz eta sailkapen algoritmo egoki bat aplikatuz sailak lortuko dira.

IX.6.1. Goranzko Sailkapen Hierarkikoa

Sailkapen automatikoko moten artean, zentruaren ingurukoa, beheranzko sailkapen hierarkikoa eta goranzko sailkapen hierarkikoa ditugu. Azken hau da garrantzitsuena duen koherentzia logikoagatik, hain zuzen.

Goranzko sailkapena egitean elementuak binaka elkartzen dira; honela, lehen etapan, bi hurbilenak elkartu ondoren m-1 elementu geratzen dira. Elkartuz lortu den elementua besteak bezala izanik, bigarren etapan, berriro bi hurbilenak elkartu ondoren m-2 elementu geratzen dira; honela jarraituz, azkenean, sail bat besterik ez da geratzen, honek elementu guztiak dituelarik.

Sortzen den zuhaitz hierarkikoan ebaketa bat eginez sailak lortzen dira.



Zuhaitz hierarkikoa

Dakusagunez, lehen etapan 6, 7 elementuak elkartu dira eta elkarketa honi, (8) adabegiaz, ordenatuen ardatzean elkartze-indize txikiena dagokio; bigarren etapan 4, (6,7) elementuak elkartu dira (9) adabegian; hirugarrenean 3,2....

(8), (9), (10)..... hierarkiaren adabegiak deitzen dira.

Elementuen arteko hurbilketak lehendik definitutako distantzia baten bidez finkatuko dira.

Goranzko sailkapen hierarkikoa irizpide desberdinen arabera egin daiteke, (saltu txikiena, batezbesteko distantzia, barizentrua...), hauen artean bariantzaren metodoa delarik garrantzitsuena.

IX.6.1.1. Bariantzaren metodoa

Bariantzaren metodoaren baldintza bakarra sailkatu nahi ditugun elementuak espazio euklidear batean aurkeztuak izan ahal izatea da. Honek zera esan nahi du: I multzoko m elementuak n aldagaien arabera edo n zutabeen arabera kontingentzi taulatan bezala, karakterizatuak daude. m elementuak $\mathbb{R}^n$ espazio bektorialeko bektoreak dira, non distantzia euklidear bat definitua baita.

Bi dimentsioko espazioan hiruki zuzen batean Pitagoras-en teorema betetzen duen edozein distantziari, distantzia euklidearra deritzo.

"Aldagai-indibiduo" taulatan ohizko distantzia euklidearra definitzen da (ikus gai honen 3. atala) eta kontingentzi taulatan distantzia euklidear berezi bat χ^2 distantzia da (ikus gai honen 4. atala), hain zuzen. Ondorioz, Goranzko Sailkapen Hierarkikoa bariantzaren metodoaz edozein kasutan aplika daiteke.

Bariantzaren metodoa ez da bakarrik distantzien kalkuluan oinarritzen, inertzien kalkuluetan ere oinarritzen da, inertzia elementuen masa bider distantzia karratuen batura bezala definiturik. Kalkuluan sartzen diren elementuen masen batura bat baldin bada, inertzia bide batez bariantza izango da.

Köening-Huygens-en teorema

Suposa dezagun sailkatu nahi ditugun N(I) hodeiaren m puntuei x_i deritzegula eta p_i berorien masak direla.

$g \in \mathbb{R}^n$ eta $y \in \mathbb{R}^n$ hodeiaren grabitate-zentrua eta hodeiaren edozein puntu izanik:

$$I_N(y) = \sum_{i=1}^{m} p_i (x_i, y)^2$$

Hau da, hodeiaren inertzia y puntuarekiko, puntuen masak bider puntu horrekiko distantzia karratuen batura da.

Köening-Huygens-en teoremaren bidez zera dakigu:

$$I_N(y) = d^2(g,y) + I_N(g)$$

hau da, $I_N(g)$ grabitate-zentruarekiko hodeiaren sakabanatzea minimoa dela.

Sortutako sailak beraien elementuetan homogenoak izatea eta sailen arteko desberdintasunak ahal den hoberen finkatuta izatea nahi dugunez, homogenotasun neurria bezala "sail barneko inertzia"-ren kontzeptua eta desberdintasun neurria bezala "sail arteko inertzia"-ren kontzeptua hartuko dira.

$N_1, N_2, \ldots N_k, \ldots N_h$; N(I) hodeiko h sail disjuntu izanik eta g_k, $P(N_k)$, N_k sailaren grabitate-zentrua eta masa izanik:

$$I_{N_k}(g_k) = \sum_{x_i \subset N_k} p_i \, d^2(x_i, g_k)$$

Hau da, N_k sailaren grabitate-zentruarekiko daukan "sail barneko inertzia" $I_{N_k}(g_k)$ balioa da.

N(I) hodeiaren g grabitate-zentruarekiko "sail arteko inertzia" hauxe izango da:

$$\sum_{k=1}^{h} p(N_k) d^2(g_k, g)$$

Köening-Huygens-en teorema aplikatuz:

$$I_N(g) = \sum_{k=1}^{h} I_{N_k}(g_k) + \sum_{k=1}^{h} p(N_k) \, d^2(g_k, g)$$

Hau da, N(I) hodeiaren inertzia, g grabitate-zentruarekiko sail barneko inertzien baturan gehi sail arteko inertzian deskonposatzen da batukorki.

I_N (g) konstante izanik, sail barneko inertzien batura minimizatzea eta sail arteko inertzia maximizatzea baliokidea da.

Algoritmoaren konkrezioa

Elkartzearen irizpidea, N(I) hodeiaren inertziaren galtzea minimoa izatea da. Honela, lehen, klaseak elementu guztiak izanik, sail barneko inertzia minimoa daukaten bi klaseak elkartuko dira.

x_i, x_j lehen fasean ditugun m klaseen bi elementu izanik, sail barneko inertzia hauxe da:

$$\Delta I_{ij} = \frac{p_i\ p_j}{p_i + p_i}\ d^2\ \left(x_i, x_j\right)$$

Honela, ΔI_{ij} minimoa daukaten bi elementuak elkartuko dira.

Sailkapenaren hurrengo klaseetan z_k eta x_i, x_j elkartzean sortu den beste edozein klaseren arteko sail barneko inertzia, azken honen grabitate-zentrua x izanik ondoko errepikapen-formulaz lortzen da.

$$\Delta\ I_{z,x} = \frac{1}{p_k + p_i + p_j}\left[\left(p_k + p_i\right)\Delta I_{ki} + \left(p_k + p_j\right)\Delta I_{kj} + p_k\ \Delta I_{ij}\right]$$

Era honetan, fase bakoitzean inertziaren galtze minimoa eragiten duten bi sailak elkartzean, hodeiarekiko sail barneko inertzia minimoa eta bide batez sail arteko inertzia maximoa edukitzea lortzen da.

Dakigunez, elementu guztiak sail batean elkartzean prozesua amaituko da.

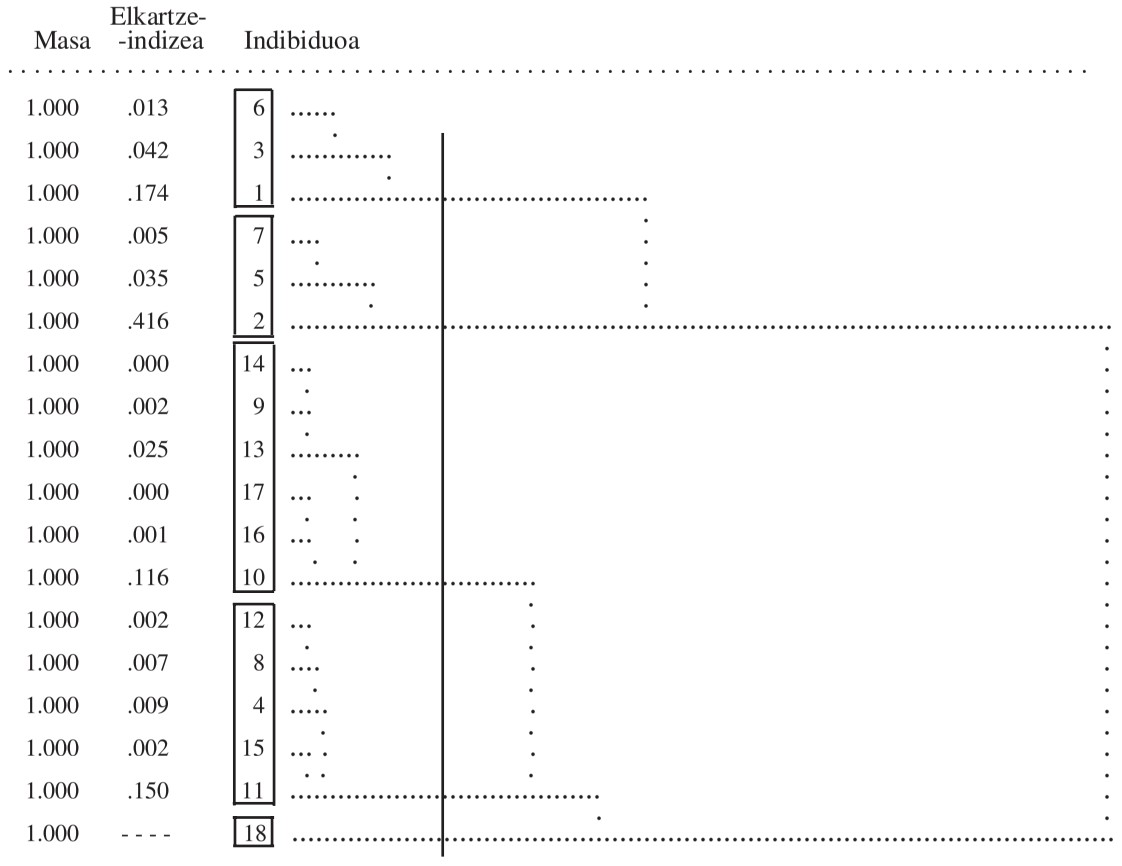
Hodeiaren partizio ezberdinak zuhaitz hierarkikoaren ebaketaz lortu ahal dira.

IX.6.2. Adibidea

Hamazortzi Hego eta Ertamerikako estatuei buruz egindako Osagai Nagusien Analisiaren azterketan (3.4 atala) hiru lehen faktoreak kontutan harturik, hau da, hauen gain estatuek dituzten koordenatuen arabera sailkapen automatikoa egiteaz ikasketa osotu nahi dugu.

Goranzko sailkapen hierarkikoa bariantzaren metodoaz aplikatu ondoren lortutako emaitzak ondokoak dira:

Zuhaitz Hierarkikoa



Elkartze-indizeak salto handia eman duelako zuhaitzaren adarrak nahiko luzatzen direnean ebaketa eginez bost sail ditugu.

Dakusagunez, zuhaitzaren oinarrian sailetan ditugun indibiduoek datu-taulan daukaten posizioaren arabera ditugu; halaber, indibiduoen masa 1 dela ikus daiteke.

Aplikatutako SPAD programen bidez sailen deskripzio zehatza eta lehen plano faktorialean estatu guztiak bere sailen zenbakiaz adierazita lortzen ditugu:

```
.....

                    FIN DE L - ETAPE     ** SEMIS **

.....

CLASSE EFFECTIF     POIDS  CONTENU (ANCIENNES CLASSES)
.....

    1          3     3.000     1     3     6
    2          3     3.000     2     5     7
    3          5     5.000     4     8    11    12    15
    4          6     6.000     9    10    13    14    16    17
    5          1     1.000    18

.....

VEN / 1 ARG / 2 MEX / 1 PAN / 3 URU / 2 BRA / 1 TXI / 2 KOL / 3 CRI / 4 GUA / 4 EKU / 3
PER / 3 PAR / 4 NIK / 4 DOE / 3 HON / 4 ELS / 4 BOL / 5
.....
```

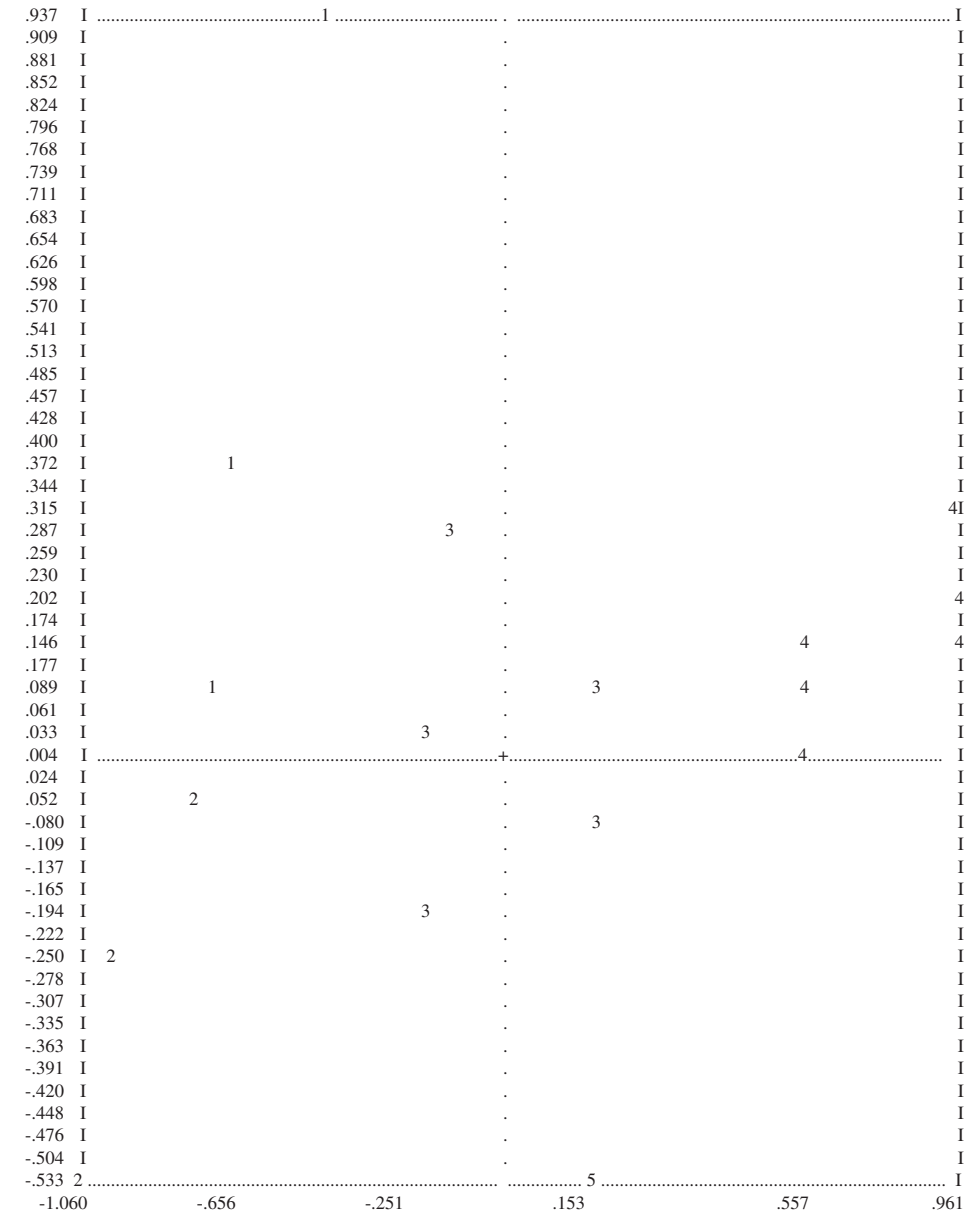
GRAPHIQUES DE DENSITE DES 18 INDIVIDUS

PLAN DE PROJECTION DES 18 POINTS SUR LES AXES 1 ET 2

AXE 1 / HORIZONTAL AXE 2 / VERTICAL

1 POINTS A PLUS DE 2.3 ECARTS-TYPES DE L-ORIGINE ONT ETE RAMENES SUR LE CADRE

1



Dakusagunez, plano faktorial nagusian, sailak oso ondo bereiziak agertzen dira. Lehen sailean eta goitik behera, Venezuela, Mexiko eta Brasil. Bigarrenean, Txile, Uruguay eta Argentina, hego-konoa hain zuzen. Hirugarrenean, Panama, Kolonbia, Peru, Ekuador eta Dominikar Errepublika. Laugarrenean C. Rica, Nikaragua, Paraguai, Guatemala, Honduras eta El Salvador. Bosgarrenean, azkenik, duen inflazioagatik hain berezia den Bolivia.