ESTATISTIKA ETA DATUEN ANALISIA

IX: Bi aldagai kuantitatiboen baterako azterketa: korrelazioa

Egilea: Josemari Sarasola

Gizapedia

gizapedia.hirusta.io

9.1 Atributu-aldagai kuantitatibo erlazio estatistikoa

9.1.1 Aldagai independentea: atributua

9.1.1.1 Eta edo korrelazio-ratioa

9.1.2 Aldagai independentea: kuantitatibo diskretua

9.1.3 Aldagai independentea: kuantitatibo jarraitua

9.2 Korrelazioa aztertzeko abiapuntua: puntu-hodeia

9.3 Korrelazio-motak

9.4 Kobariantza

9.5 Pearson-en korrelazio koefiziente lineala

9.6 Aldagai dikotomiko-aldagai kuantitatibo korrelazioa

9.7 Aldagai dikotomiko-aldagai dikotomiko korrelazioa

9.8 Korrelazio partziala

9.9 Sasiko korrelazioa

9.10 Ariketak

9. gaia: Bi aldagai kuantitatiboen baterako azterketa: korrelazioa

Aurreko ikasgaian bi aldagai kualitatiboen arteko asoziazioa ikasita, ikasgai honetan bi aldagai kuantitatiboen arteko erlazio estatistikoa, **korrelazioa** alegia, ikasi behar dugu. Hori baino lehen, ordea, tarteko egoera bat landuko dugu: atributu baten eta aldagai kuantitatibo baten arteko erlazioa alegia.

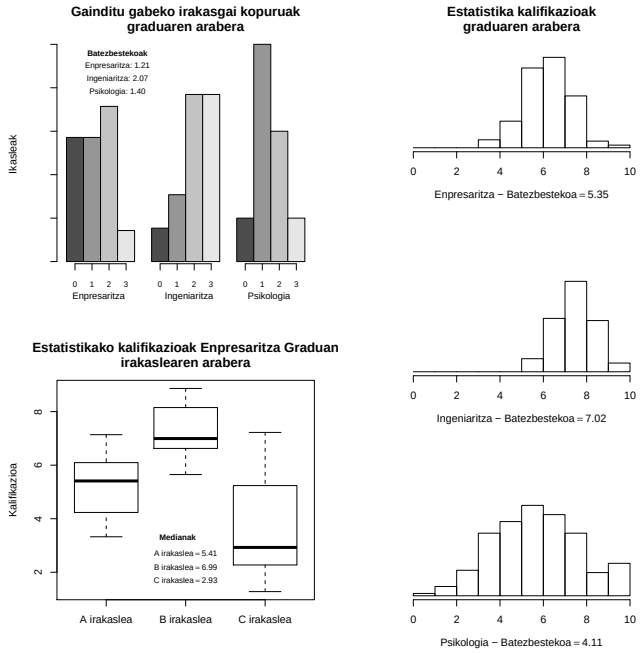
9.1 Atributu-aldagai kuantitatibo erlazio estatistikoa

Aldagai kualitatibo baten eta kuantitatibo baten arteko erlazio estatistikoa aztertzeko lehen pausoa **aldagai independentea eta dependentea** zehaztea da. Ondoren, prozedura orokorra aldagai dependentearen datuak aldagai independentearen kategoria, balio edo balio-tarteen arabera bereizi eta datu-azpimultzo bakoitza bere aldetik aztertu eta azterketa horiek elkarrekin alderatzea da.

9.1.1 Aldagai independentea: atributua

Kasu honetan, atributuaren kategoriaren arabera datuak bereizi eta sortutako datu-azpimultzo kuantitatiboak (beste aldagai kuantitatiboari buruzkoak) banan-banan aztertu behar dira, horretarako diagrama egokiak erabiliz (puntu-diagramak, histogramak, kaxa-diagramak) eta estatistiko deskribatzaileak erabiliz (batezbesteko aritmetikoa, desbideratze estandarra eta abar).

Irudiotan adibide batzuk azaltzen dira:



Irudia 9.1: **Atributua da aldagai independentea**. *(Goian ezkerrean)* Gradua eta gainditu ez diren irakasgaien kopurua jaso dira ikasle batzuengan. Aldagai independentea gradua da, eta horren arabera banatzen dira datuak azterketa egiteko. Batezbestekoei erreparatuz, enpresaritzan gutxiago eta ingeniaritzan gehiago suspenditzen da. *(Behean ezkerrean)* Irakaslea eta estatistikan izandako kalifikazioa jaso dira Enpresaritzako ikasle batzuen artean. Aldagai independentea irakaslea da, irakaslearen arabera izan baitaitezke ezberdinak notak. B irakaslearekin nota handiagoa lortzen da orokorrean. *(Eskuinean)* Ikasle batzuen artean estatistika nota eta gradua jaso dira. Graduaren arabera lortzen da estatistikan nota handiagoa edo txikiagoa, beraz gradua da aldagai independentea. Ingeniaritzan lortzen da orokorrean nota handiena.

9.1.1.1 Eta edo korrelazio-ratioa

Korrelazio-ratioak, eta (η) ere deituak, aldagai kualitatiboak aldagai kuantitatiboak zenbateraino azaldu edo esplikatzen duen neurtzen du. Honela kalkulatzen da:

$$\eta^2 = \frac{\sum_x n_x(\overline{y}_x - \overline{y})^2}{\sum_{x,i}(y_{xi} - \overline{y})^2}$$

non,

- $\overline{y}$ datu guztien batezbestekoa den,
- $\overline{y}_x$ x kategoriako batezbestekoa den,
- y_{xi} datu bakoitza den.

η-ren balio zehatza asoziazio-neurrien balioa bezalatsu interpretatzen da.

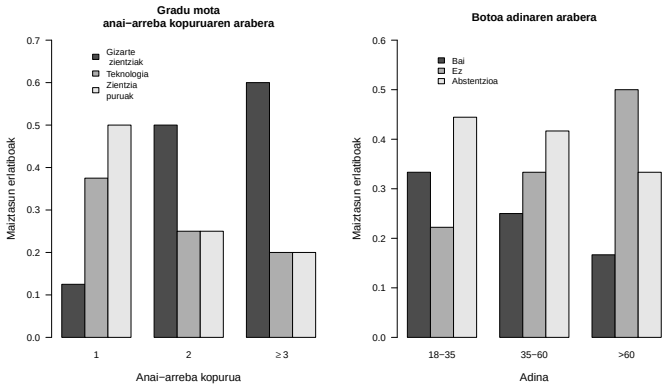
Atributuaren kategoriak bi bakarrik direnean, Cohen-en d erabil daiteke aukeran, baina horretarako gogoratu behar da bi kategorietako datuen bariantzak berdinak izan behar direla.

9.1.2 Aldagai independentea: kuantitatibo diskretua

Adibidez, pertsona batzuengan anai-arreba kopurua eta aukeratutako karrera jaso direnean, aldagai independentea anai-arreba kopurua da. Hala, bi aldagaien arteko erlazio estatistikoa aztertzeko, aukeratutako gradu-motari buruzko datuak anai-arreba kopuruaren arabera sailkatuko ditugu. Ondoren, gradu-motari buruzko datu-azpimultzo bakoitza bere aldetik aztertu (maiztasun-taulak eta barra-diagramak baliatuz esaterako) eta emaitzak (gradu bakoitzaren portzentajeak) elkarrekin alderatuko dira.

9.1.3 Aldagai independentea: kuantitatibo jarraitua

Adibidez, pertsona batzuengan adina eta bozketa batean emandako botoa (bai-ez-abstentzioa) jaso direnean, aldagai independentea adina da. Hala, botoa adinaren arabera aztertuko dugu. Adinak balio ezberdin asko hartzen dituenez, adin-tarteak osatu eta tarte horien arabera botoari buruzko datu-azpimultzoak izango ditugu. Hain zuzen, aldagai kuantitatiboak balio asko hartzen dituenean, beste aldagaiari buruzko datu-azpimultzoak sortzeko modu bakarra aldagai kuantitatiboan tarteak osatzea da. Adibidearekin jarraituz, adin-tarte bakoitzean sektore ekonomikoari buruzko portzentajeak kalkulatu eta beste tarteetakoekin alderatu ditugu. Barra-diagramak ere era daitezke alderatzeko.



Irudia 9.2: **Aldagai independentea kuantitatiboa da**. *(Ezkerrean)* Anai-arreba bakarren kasuan, zientzia eta teknologia dira gehien aukeratzen direnak. Anai-arreba kopuru handiko familietan, berriz, gizarte-zientziak nagusitzen dira. *(Eskuinean)* Ikusten denez, zenbat eta adin handiagoa, baiezko botoa emateko joera handiagoa da.

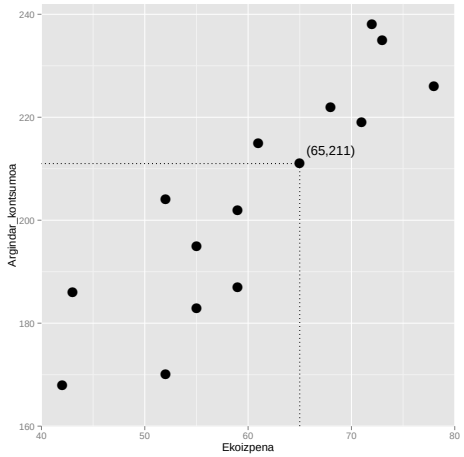
9.2 Korrelazioa aztertzeko abiapuntua: puntu-hodeia

Hemendik aurrera **korrelazioa** soilik, bi aldagai kuantitatiboen arteko erlazio estatistikoa alegia, aztertu behar dugu ikasgaian zehar. Horretarako abiapuntua bi aldagaien datuak lotzen dituen **puntu-hodeia** edo **sakabanatze-diagrama** izenekoa da (ingelesez, *scatter plot*, datu guztiak kartesiar diagrama batean iruditzen dituena besterik ez dena. Puntu-hodeia aztertuz korrelazioaren nondik norakoa azter daiteke.

Adibidea: Lantegi batean argindar-kontsumoa eta ekoizpena jaso dira 15 egunetan zehar. Honakoak dira datuak:

Ekoizpena	65	72	59	68	52	55	71	73
Argindar-kontsumoa	211	238	187	222	204	195	219	235
Ekoizpena	43	59	78	61	52	55	42	
Argindar-kontsumoa	186	202	226	215	170	183	168	

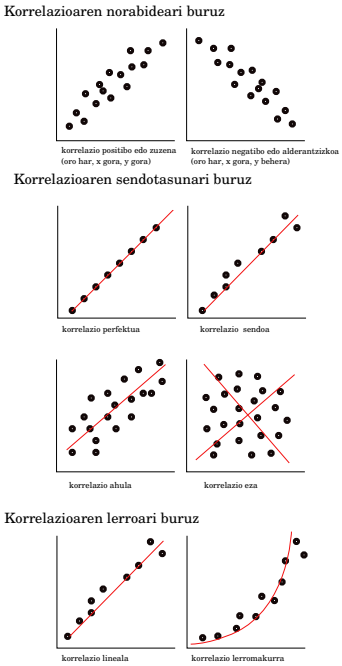
Puntu-hodeia eratu eta interpretatu behar da.



Irudia 9.3: **Puntu-hodeia**. Diagramaren eraketa erakusteko, lehen datu-bikotearen koordenatuak zehaztu dira. Ikusten denez, *oro har*, zenbat eta ekoizpen handiagoa, argindar-kontsumoa ere hainbat eta handiagoa da.

9.3 Korrelazio-motak

[22]r0.5



Norabidearen arabera,

- **korrelazio positibo** edo **zuzena**, aldagai bat gehitzean, orokorrean beste aldagaia ere gehitu egiten denean;
- **korrelazio negatibo** edo **alderantzizkoa**, aldagai bat gehitzean, orokorrean beste aldagaia murriztu egiten denean.

Sendotasunaren arabera,

- **korrelazio perfektua**, puntuak lerro berean daudenean;
- **korrelazio sendoa**, aldagaien arteko korrelazioa argia eta estua denean;
- **korrelazio ahula**, aldagaien arteko korrelazioa lausoa denean;
- **korrelazio hutsa**, aldagaien artean batere korrelazioa badaezpadakoa denean.

Lotura funtzionalaren arabera,

- **korrelazio lineala**, aldagaien arteko korrelazioa zuzen baten araberakoa denean gutxi gorabehera;
- **korrelazio lerromakurra**, aldagaien arteko korrelazioa kurba baten araberakoa denean gutxi gorabehera.

9.4 Kobariantza

Kobariantza *korrelazio linealaren norabidea* neurtzen duen koefiziente bat da. Nabarmendu behar dira *korrelazio lineala* soilik neurtzen duela, eta horren *norabidea*. Ez du korrelazioaren sendotasunari buruz inongo informaziorik ematen. Bi formula hauekin kalkula daiteke:

$$s_{xy} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{\sum_i x_i y_i}{n} - \overline{xy}$$

Lehen formula kobariantzaren definizioari eta jatorrizko eraketari dagokio. Bigarrenarekin aiseago burutzen dira kalkuluak.

Kobariantzak edozein balio har dezake, positibo zein negatibo, inongo mugarik gabe. Honela interpretatzen da: kobariantza positiboa bada, korrelazio positibo edo zuzena da; negatiboa bada, korrelazio negatiboa edo alderantzizkoa da. Arestian adierazi bezala, ez du korrelazioaren sendotasunari buruz inongo informaziorik ematen.

Beste alde batetik, kobariantzaren unitateak alagaien unitateen biderkadura da; adibidez, altuerak cm-tan eta pisuak kg-tan jasota, bi aldagaien arteko kobariantzaren unitateak $cm \times kg$ dira.

9.5 Pearson-en korrelazio koefiziente lineala

Kobariantzak sendotasunari buruzko informazioa ez ematearen arrazoia unitateekin du zerikusirik: adibidez, altueraren (cm) eta pisuaren (kg) arteko kobariantza kalkulatzean, altuera cm-tan eman ordez m-tan emango bagenu kobariantza berria cm kalkulatutakoa zati 100 izango lizateke. Noski, horregatik ezin dugu esan korrelazio txikiagoa denik, unitatea bakarrik aldatu diegunez, datuak berdinak direlako.

Sendotasuna neurtzeko egokia izateko, beraz, unitatea ezabatu beharko diegu datuei, datuak dimentsiogabetu alegia, eta hori jada ikasita daukagun estandarketa delakoaz egiten dugu (gogoratu $z = (x - \overline{x})/s$). Beraz, kobariantza x eta y datuekin kalkulatu ordez, z_x eta z_y datuekin eman beharko da:

$$s_{z_x z_y} = \frac{\sum z_x z_y}{n} - \overline{z_x} \cdot \overline{z_y}$$

Lehenbizi, aldagai estandartuen batezbestekoa 0 dela frogatuko dugu:

$$\overline{z_x} = \frac{\sum \left(\frac{x_i - \overline{x}}{s_x}\right)}{n} = \frac{1}{s_x} \frac{\sum (x_i - \overline{x})}{n} = \frac{1}{s_x} \left(\frac{\sum x_i}{n} - \frac{n\overline{x}}{n} \right) = \frac{1}{s_x} (\overline{x} - \overline{x}) = 0$$

Horrela,

$$s_{z_x z_y} = \frac{\sum z_x z_y}{n} = \frac{\sum \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n} = \frac{1}{s_x s_y} \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{s_{xy}}{s_x s_y}$$

Azken adierazpenari Pearson-en korrelazio-koefiziente lineala deritzo, eta honela adierazten da:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

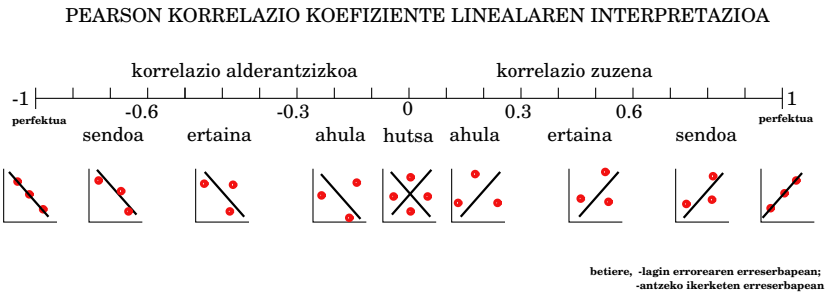
Koefizienteak [-1,1] tarteko balioak hartzen ditu, korrelazio lineala soilik neurtzen du (kobariantzak bezalaxe), eta honela interpretatzen da:

- **norabideari buruz**,
  - $r_{xy} > 0$ badugu, korrelazioa positiboa edo zuzena da;
  - $r_{xy} < 0$ badugu, korrelazioa negatiboa edo alderantzizkoa da;
- **sendotasunari buruz**,
  - $|r_{xy}| < 0.3$ badugu, korrelazioa ahula dela esango dugu;
  - $0.3 < |r_{xy}| < 0.6$ badugu, korrelazioa ertaina dela;

– $|r_{xy}| > 0.6$ badugu, korrelazioa sendoa dela;

– $|r_{xy}| = 1$ badugu, korrelazioa perfektoa da (puntu guztiak zuzen batean lerrokatuta daude).

Kasu guztietan, intepretazioa antzeko ikerketen emaitzen eta lagin-errorearen erreserbapean izango da.



Irudia 9.4: Korrelazioa interpretatzeko erregela.

9.6 Aldagai dikotomiko-aldagai kuantitatibo korrelazioa

Pearsonen korrelazio-koefizientea kalkulatu ahal izateko, bi aldagaiak kuantitatiboak izan behar dira. Aldagai kualitatiboren bat tartean dagoenean, kalkuluak ezin dira noski egin. Aldagaia *dikotomikoa* denean, bi kategoria bakarrik hartzen dituenean alegia (adibidez, sexuaren kasuan, gizon/emakume; azterketa bat gainditu den, bai/ez) ordea, ahal da korrelazio-koefiziente lineala kalkulatu, aldagai dikotomikoaren bi kategoriei 0 eta 1 esleituz, hurrenez hurren. Interpretazioa egiterakoan, tentuz ibili behar da, eta 0/1 balioak behar bezala deskodetu behar dira.

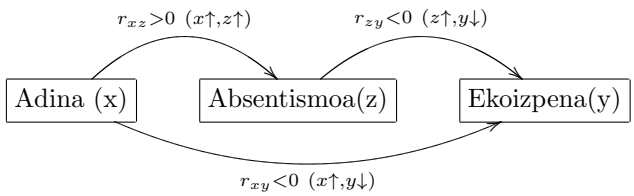
9.7 Aldagai dikotomiko-aldagai dikotomiko korrelazioa

Bi aldagaiak dikotomikoak direnean ere kalkula daiteke korrelazio-koefiziente lineala. Nahikoa da bi aldagai dikotomikoetako bina kategoriei 0/1 balioak esleitzea. Interpretazioa 0/1 balioak deskodetuz egiten da.

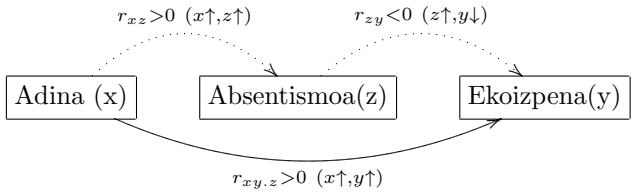
9.8 Korrelazio partziala

Demagun lantegi bateko langileen adinak, azken urtean bakoitzak izan dituen absentismo-egunak eta ekoizpena jaso direla. Hipotesi teorikoak dioenez, langilea zenbat eta zaharragoa (adina zenbat eta handiagoa), ekoizpena hainbat eta handiagoa da. Adinaren eta ekoizpenaren arteko korrelazio-koefizientea kalkulaturik, ordea, negatiboa ematen du. Zer dela-eta suertatzen da uste denaren aurkakoa? Tarteko aldagai bat dago, absentismoa alegia, emaitzak distortsionatu egiten dituena. Hain zuzen, adinean gora, absentismoak ere gora egiten du, eta horrek ekoizpena behera dakar. Beraz, absentismoa da, horren tarteko eraginagatik, ekoizpena behera ekarri eta adinean gora ekoizpenak behera egiten duelako itxurazko korrelazioa sorrarazten duena. Hori saihesteko, tarteko aldagaiaren eragina saihestu edo ezabatu behar da, kasu honetan absentismoarema. Hori korrelazio-koefiziente partzialaren bitartez egiten da. Adina x, absentismoa z eta ekoizpena y izanik, honela kalkulatzen da x eta y aldagaien arteko korrelazio-koefiziente partziala, z aldagairen efektua baztertuz:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$



Irudia 9.5: Absentismoaren tarteko eraginagatik (goiko geziak), adinaren eta ekoizpenaren arteko korrelazioa negatibo suertatzen da (beheko gezia), aldez aurretik uste denaren bestera, adinak lanpostuan esperientzia handiagoa dakarrelako.



Irudia 9.6: Oztopoa gainditzeko, absentismoaren eragina baztertu behar da (goiko geziak), korrelazio partzialeko koefiziente baten bitartez (beheko gezia), orokorrean korrelazio positiboa emango duena, uste denarekin bat.

9.9 Sasiko korrelazioa

Sasiko korrelazioa edo **korrelazio faltsua** elkarrekin zerikusi handirik ez duten aldagaiak uztartean gerta daiteke. Adibidez, biztanleria nahiko konstantea duen herrialde bateko urteko ogi-kontsumoaren eta hurrengo urteko jaiotza-kopuruaren arteko korrelazio-koefizientea kalkulaturik, bi aldagaien arteko korrelazio-koefizientea positibo eta sendoa suertaturik, ezin da ondorioztatu, ogia afrodisikoa denik edota bikoteen ugalkortasuna gehitzen duenik, bi aldagaiak horrela lotzeko fundamenturik ez dagoelako; bi aldagai horien arteko korrelazioa sasikoa dela esango dugu orduan. Beraz, aldagaien arteko korrelazioa aztertzerakoan, aldagai horiek logikaz loturik egon behar dira, teoria baten baitan, emaitzak adierazgarriak izango badira.

9.10 Ariketak

1. A, B eta C geletan kalifikazio hauek jaso dira:

A	8	5	6	9
B	6	6	7	7
C	5	6	5	-

   Eta korrelazio-ratioa kalkulatuta, gelaren eragina neurtu ezazu kalifikazioetan.

2. Ikasle zenbaiten gainean, matematika nota, bigarren hizkuntzako nota, 200 m korritzeko behar duten denbora (segundutan) eta egunean zehar telefono mugikorrarekin zenbat denbora ibili diren (minututan):

Matematika	5.2	5.2	8.3	8.9	4.3	7.4	7.8	9.0	5.6
Hizkuntza	3.4	5.4	7.8	9.2	2.1	6.8	8.3	8.7	4.5
Lasterketa	36	77	58	42	91	74	24	35	56
Telefonoa	45	54	12	15	66	28	22	10	39

   Matematika notaren eta beste hiru aldagaien arteko puntu hodeiak marraztu eta interpretatu behar dira.

3. Enpresa zenbaitetan I+G alorrean izandako gastu eta mozkin totalen portzentajea jaso da fakturazioari buruz:

I+G	2	3	1	4	6
Mozkinak	15	20	12	24	22

   (a) Bi aldagaien arteko kobariantza kalkulatu eta interpretatu, formula erosoa nahiz jatorrizkoa baliatuz.

   (b) Bi aldagaien arteko korrelazio-koefiziente lineala kalkulatu eta interpretatu. Esan al daiteke ziurtasun handiz I+G gastuak mozkinak gehitzen dituela?

4. Enpresa batean asteko ekoizpenari eta unitate kostuari buruzko datuak jaso dira aste zenbaitetan zehar:

Astea	1	2	3	4	5	6	7	8	9	10	11	12
Ekoizpena	10	12	15	18	22	27	33	39	42	48	56	68
Unitateko kostua	99	58	48	41	38	36	35	35	34	34	34	34

   (a) Kobariantza kalkulatu eta interpretatu.

   (b) Korrelazio-koefiziente lineala kalkulatu eta interpretatu.

   (c) Bi aldagaien artean nolako erlazio estatistikoa dago: lineala ala lerromakurra? Puntu hodeia marraztu galdera erantzuteko.

5. Ikasle zenbaiten sexua kontuan harturik, froga batean izandako kalifikazioak jaso dira:

Sexua	g	e	g	e	g	e	e
Kalifikazioa	6.7	8.5	6.9	7.2	8.0	8.6	9.2

   Korrelazio koefiziente lineala kalkulatu eta interpretatu. Zenbateraino du eragina sexuak kalifikazioetan?

6. Test bateko puntuazio totala eta galdera jakin bat ongi erantzun duten galdetu zaie ikasle zenbaiti (galdera, z: zuzen, o: oker):

Puntuazioa	87	46	65	72	82	61	39	42
Galdera	z	o	z	o	o	z	z	z

   Galdera hori baztertzea komenigarria den azter ezazu.

7. Botika berri bat asmatu ondoren, urtebetez frogatu zen gaixotasun bat pairatzen zuten pertsonen zenbaiten artean. Beste batzuek ohiko tratamenduarekin jarraitu zuten. Urtebete pasa ondoren, gaixotasun-egoera arindu den jaso zen. Emaitzak honako hauek dira:

Botika berria?	b	b	b	b	b	e	e	e	e	e
Gaixotasuna arindu?	b	e	b	b	b	e	e	e	e	b

   Bi aldagaien arteko korrelazio koefiziente lineala baliatuz, erabaki botika eraginkorra izan daitekeen.

8. Test bat gainditu duten ala ez eta testeko galdera jakin bat ongi erantzun duten ala ez jaso da azterketa bat egin duten pertsona zenbaiten artean. Emaitzak hauek dira:

	Galdera ongi erantzun da?		Guztira
Testa gainditu?	Bai	Ez	
Bai	36	16	52
Ez	12	54	66
Guztira	48	70	118

   Bi aldagaien arteko korrelazio koefiziente lineala baliatuz, azter ezazu galderaren garrantzia testaren kalifikazioa zehazterakoan.

9. Administrazio publikoko oposizioetan lehiakideen artean, sexua (g: gizon,0; e:emakume,1), maila ekonomikoa (b: baxua, 0; a:altua, 1), lorturiko nota eta oposizioa prestatzeko bi urteko prestakuntza intentsiboa eskaintzen duen akademia batera (bai, 1; ez, 0) joan diren jaso dira eta aldagai guztien arteko korrelazio-koefiziente linealak kalkulatu:

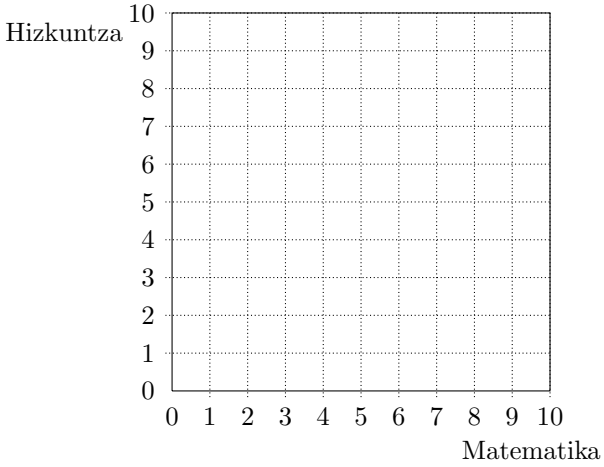
r_{xy}	Sexua	Maila ekon.	Nota	Akademia
Sexua	1	-0.258	-0.060	0
Maila ekon.	-0.258	1	0.669	0.774
Nota	-0.060	0.669	1	0.904
Akademia	0	0.774	0.904	1

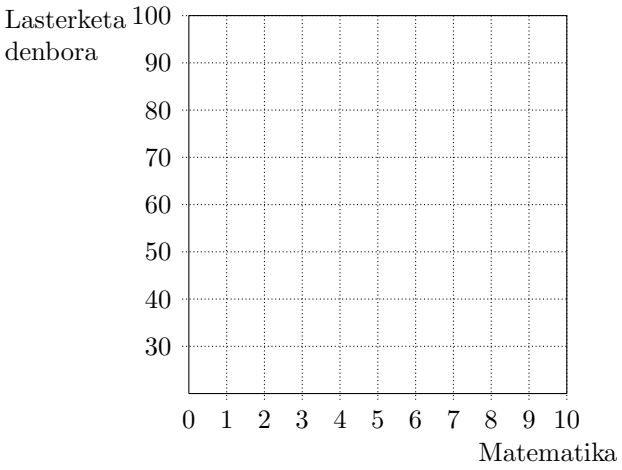
   (a) Azaldu aurreko korrelazio-matrizearen ezaugarriak.

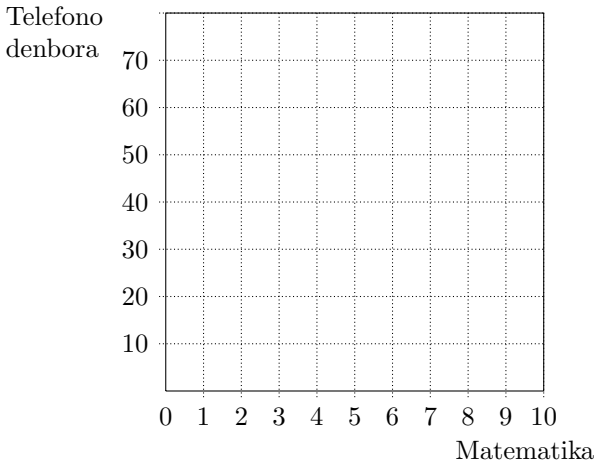
   (b) Aldagai horien arteko korrelazioari buruz hipotesiak planteatu, eta hasieran batean itxuraz betetzen ez diren kasuetan, korrelazio partzialaren kontzeptua baliatu hipotesia egiaztatzeko.

Ebazpenak

(2) ariketa







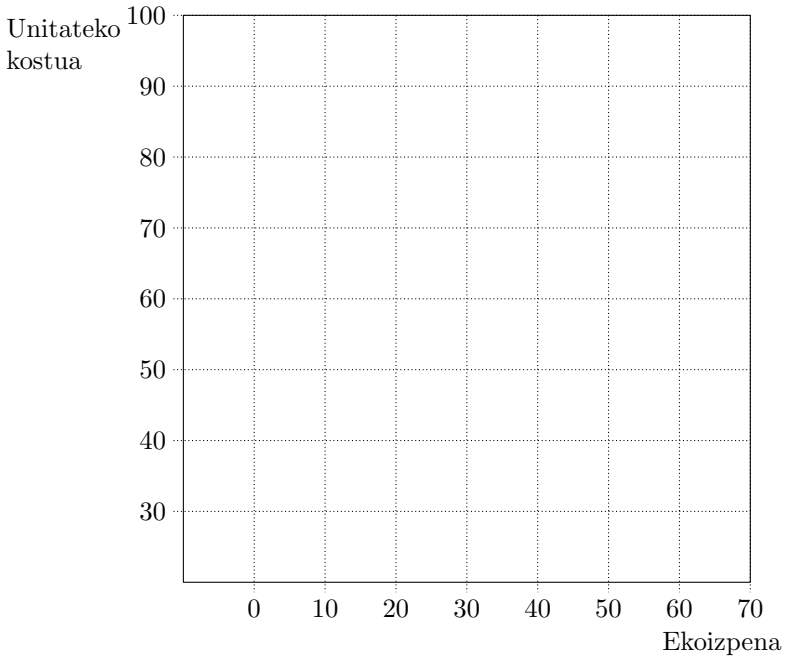
(3) ariketa

x_i (I+G)	y_i (Mozkinak)	$x_i y_i$	x_i^2	y_i^2
2	15			
3	20			
1	12			
4	24			
6	22			

x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x}) \times (y_i - \overline{y})$
2	15			
3	20			
1	12			
4	24			
6	22			

(4) ariketa

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
10	104			
12	58			
15	48			
18	41			
22	38			
27	36			
33	35			
39	35			
42	34			
48	34			
56	34			
68	34			


Unitateko kostua
100
90
80
70
60
50
40
30
0
10
20
30
40
50
60
70
Ekoizpena