ESTATISTIKA ETA DATUEN ANALISIA

IV: Aldagai bakunaren deskribapena: sakabanatzea

Egilea: Josemari Sarasola

Gizapedia

gizapedia.hirusta.io

4.1 Sakabanatzearen kontzeptua

4.2 Sakabanatze-neurri absolutuak

4.2.1 Ibiltartea

4.2.2 Kuartil arteko ibiltartea

4.2.3 Desbideratze estandarra eta bariantza

4.2.3.1 Bariantza

4.2.3.2 Populazio-bariantza eta lagin-bariantza

4.2.3.3 Kalkulua tartekako datuekin

4.2.3.4 Desbideratze absolutuen mediana

4.3 Sakabanatze-neurri erlatiboak

4.4 Estandarketa

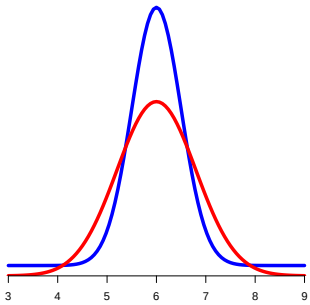
4.5 Efektuaren tamaina

4.6 Ariketak

4. gaia: Aldagai bakunaren deskribapena: sakabanatzea

4.1 Sakabanatzearen kontzeptua

Datu-multzo kuantitatiboen ezaugarri bakarra ez da zentroa. Beste ezaugarri garrantzitsu bat **sakabanatzea** edo *aldakortasuna* da, datuak beraien artean nahiz zentrotik zenbateraino desbideratzen diren jasotzen duena. Ikasiko ditugun sakabanatze-neurrietan ikasiko dugunez, sakabanatzea neurtzeko irizpide nagusia datuen arteko distantzia edo datu guztietatik zentrora dagoen distantzia da.



Irudia 4.1: Azterketa bateko puntuazioak, gizonezko (gorriz) nahiz emakumezkoetarako (urdinez). Bi sexuak 6 puntuazioaren inguruan biltzen dira, eta ondorioz zentro berdina dute, baina gizonezkoen puntuazioak sakabanatuagoak dira.

4.2 Sakabanatze-neurri absolutuak

4.2.1 Ibiltartea

Ibiltartea (ingelesez, *range*) datu txikienetik handienera dagoen distantzia da:

$$R = x_{max} - x_{min}$$

Neurri hau zenbat eta handiagoa izan, orduan eta sakabanatzea handiagoa dagoela ondorioztatuko da. Neurri honek bi oztopo ditu sakabanatzea neurtzeko:

- ez da sendoa, hau da, datu atipikoek nabarmen eragiten dute;
- ez du datuetan jasotzen den informazio guztia biltzen, hots, ez ditu datu guztiak erabiltzen.

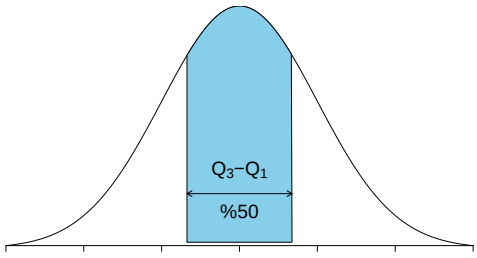
Baina hain zuzen ere, sendoa ez izateagatik datu atipikoak aurtkitzeko tresna gisa erabiltzen da maiz, bereziki kalitate kontrolean.

4.2.2 Kuartil arteko ibiltartea

Sakabanatze-neurri sendo moduan **kuartil arteko ibiltartea** (ingelesez, *interquartile range*) erabiltzen da. Muturreko datuen arteko distantziaren ordez, erdian dauden datuen %50en arteko distantzia hartzen du kontuan:

$$IQR = Q_3 - Q_1$$

Zenbat eta handiagoa izan, orduan eta sakabanatze handiagoa du datu-multzoak. Eragozpenik badu: ez du kontuan hartzen datuetan jasotzen den informazio guztia.



4.2.3 Desbideratze estandarra eta bariantza

Desbideratze estandarra datu bakoitzetik batezbesteko aritmetiko sinplera duen $(x_i - \overline{x})$ distantzian oinarritzen da. Datu guztietarako distantzia horiek kalkulatu, eta batzuk positiboak eta batzuk negatiboak izango direnez, horien batezbesteko kuadratikoa kalkulatzen du:

$$s_x = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n}}$$

Horrela, desbideratze estandarrak *datu bakoitza batezbesteko aritmetikotik batezbestez zenbat desbideratzen den* adierazten du. Horrela, zenbat eta handiagoa izan, orduan eta sakabanatze handiagoa dagoela adierazten du. Aipatu behar da, bestalde, *beti balio positiboak* hartzen dituela (0 ere izan daiteke, hain zuzen datu guztiak berdinak direnean), eta aldagaiaren unitatetan neurtzen dela (datuak minututan badira, desbideratzea ere minututan izango da).

Badu eragozpen bat: ez da sendoa. Baina abantailarik ere badu: datu guztiak hartzen ditu kontuan.

Aurreko formula ez da oso erosoa kalkuluak eskuz egiteko. Baina formula hori garatzen bada, beste formula erosoago honetara heltzen gara:

$$s_x = \sqrt{\frac{\sum_i x_i^2}{n} - \overline{x}^2}$$

Adibidea: Ibilbide bat egiteko denbora hauek jaso dira (min):

22-25-28-26-24

Kalkulatu desbideratze estandarra eta interpretatu.

Formulan ikusten denez, desbideratzen estandarra kalkulatzeko lehen pausoa batezbesteko aritmetiko sinplea kalkulatzea da. Ondoren, datuen karratuen batura kalkulatu behar da. Horrekin, aski da emaitzak formulan txertatu eta kalkulua egitea. Zutabe-formatoa da egokiena kalkulu guztiak egiteko:

x	x^2
22	484
25	625
28	784
26	676
24	576
125	3145

$$\overline{x} = \frac{125}{5} = 25 \ ; \ s_x = \sqrt{\frac{3145}{5} - 25^2} = 2$$

Beraz, denbora-datu bakoitza 2 *min* desbideratzen da batezbestez 25 *min*-ko batezbestekotik.

4.2.3.1 Bariantza

Bariantza desbideratze estandarraren karratua besterik ez da:

$$s_x^2 = \frac{\sum_i (x_i - \overline{x})^2}{n} = \frac{\sum_i x_i^2}{n} - \overline{x}^2$$

Sakabanatze-neurri bezala maiz erabiltzen da, eta zenbat eta handiagoa izan, orduan eta sakabanatze handiagoa dago.

4.2.3.2 Populazio-bariantza eta lagin-bariantza

Arestiko formulak **populazio-bariantza** izenekoari dagozkio, eta jasotako datuek ikertu nahi dugun multzo osoa edo populazioa osatzen dutenean erabiltzen da; edota lagin errorea neurtzea beharrezkotzat hartzen ez denean.

Jasotako datuek populazioaren *lagin bat* osatzen dutela kontsideratu eta populazio osoko bariantza zenbatetsi edo estimatzen denean berriz, arestiko formulek gehiegiz zenbatetsi edo estimatzen dute populazioko bariantza. Zenbatespen hori zuzentzeko, **lagin-bariantza** edo *bariantza zuzendua* erabiltzen da:

$$\hat{s}_x^2 = \frac{\sum_i (x_i - \overline{x})^2}{n-1}$$

Hortik, *desbideratze zuzendua* deitzen dena eratortzen da:

$$\hat{s}_x = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n-1}}$$

Estatistika-programa informatikoek bariantza eta desbideratzen zuzenduak kalkulatzen dituzte besterik adierazi ezean. Hori ez da oztopo, populazio- eta lagin-bariantzaren artean erlazio hau betetzen baita:

$$\hat{s}_x^2 = \frac{n}{n-1} s_x^2$$

Nolanahi ere, lagin tamaina oso handia denean, lagin-bariantzaren eta populazio-bariantzaren arteko diferentzia oso txikia da.

4.2.3.3 Kalkulua tartekako datuekin

Datuak tartekako formatoan daudenean, bariantzaren (eta desbideratze estandarraren) kalkulua ohi bezala egiten da: erreferentziazko baliotzat tarteko erdipuntua hartzen da, eta hortik aurrera tarte bakoitzean zenbat elementu dauden hartu behar da kontuan.

Adibidea: Lantegi batean langileen adinak jaso dira:

Adina	Langileak (n)
15-20	2
20-25	7
25-30	8
30-35	4
35-40	1
	22

Kalkulatu langileen adinen bariantza.

Lehen pausoa batezbesteko aritmetiko sinplea kalkulatzea da, horretarako nx zutabea osatuz. Datu karratuen batura kalkulatzeko nx^2 zutabea eratzen da (kontuan hartuz x bakoitza n aldiz errepikatzen dela).

Adina	Langileak (n)	x	nx	nx^2
15-20	2	17.5	35	612.5
20-25	7	22.5	157.5	3543.75
25-30	8	27.5	220	6050
30-35	4	32.5	130	4225
35-40	1	37.5	37.5	1406.25
	22		580	15837.5

$$\overline{x} = \frac{580}{22} = 26.36$$

$$s_x^2 = \frac{15837.5}{22} - 26.36^2 = 25 \rightarrow s_x = 5.03$$

Beraz, batezbestez adin bakoitza 5.03 urte desbideratzen da 26.36ko batezbestekotik.

4.2.4 Desbideratze absolutuen mediana

Desbideratze absolutuen mediana (DAME, ingelesez MAD, *Median of Absolute Deviations*) datu guztiak kontuan hartzen dituen sakabanatze-neurri bat da, sendoa ere badena. Medianarako desbideratzeen balio absolutuen mediana da:

$$DAME = Me[|x_i - Me|]$$

Adibidea: Ibilbide bat egiteko denbora hauek jaso dira (min):

22-25-28-26-24

Kalkulatu desbideratze absolutuen mediana.

Mediana kalkulatzen da lehenbizi: 25.

Medianarako desbideratze absolutuak kalkulatzen dira: 3-0-3-1-1.

Desbideratze absolutuak ordenaturik (0-1-1-3-3), horien mediana 1 da. Beraz,

$$DAME = 1\ min$$

4.3 Sakabanatze-neurri erlatiboak

Aurreko atalean ikasitako sakabanatze-neurri absolutuak ez dira egokiak datumultzoak alderatzeko sakabanatzeari buruz. Ikus dezagun zergatik adibide batez.

Haiti eta AEBetako errentak jaso dira hainbat familien artean eta emaitza hauek eskuratu dira batezbesteko errentari buruz eta errentaren desbideratze estandarrari buruz (informazioa alegiazkoa da eta dolarretan ematen da):

Herrialdea	$\overline{x}$	s_x
Haiti	100	10
AEB	10.000	10

Desbideratze estandarrari erreparatuta, *badirudi* errentaren sakabanatzea berdina dela bi herrialdeetan, baina hori ez da horrela, ezbaitira berdinak 10 dolar desbideratzea 100 dolarreko batezbesteko batetik eta 10.000 dolarreko batezbesteko batetik.

Beraz, sakabanatze-neurri absolutu bat beste datu-multzoen sakabanatzearekin

alderatzeko egokia izan dadin, batezbesteko bati egin behar zaio erreferentzia. Eta horretarako sakabanatze-neurri erlatiboak ditugu:

- **aldakortasun-koefizientea**: $A_X = \frac{s_x}{\overline{x}}$
- **kuartil arteko ibiltarte erlatiboa**: $RIQR = \frac{Q_3 - Q_1}{Me}$
- **DAME erlatiboa**: $DAMEerl = \frac{DAME}{Me}$

Neurri horiek ehunekotan eman ohi dira (bider ehun eginda) eta sakabanatze-neurriaren balioa adierazten dute zentro-neurri bati buruz. Dimentsiogabeak dira, hau da, ez dute unitaterik, eta horregatik hain zuzen erabil daitezke datu-multzo ezberdinen sakabanatzeak alderatzeko. Zenbat eta handiagoak izan, sakabanatzea orduan eta handiagoa da, baina hala ere ezin da zehaztu sakabanatzea handitzat har daitekeen balio bat, sakabanatzea ezaugarri erlatiboa baita: beste datu-multzo batekoa baino handiagoa edo txikiagoa izango da beti.

Alderatu beharreko datu-multzoetako batezbestekoak edo erabiltzen den beste zentro-neurriak berdinak (edo berdintsuak) direnean soilik onar daiteke sakabanatze-neurri absolutuak erabiltzea, erlatiboen ordez.

4.4 Estandarketa

Estandarketa datuak transformatu egiten dituen aldagai-aldaketa bat da, emaitza moduan *datu estandartu* bat ematen duena (x_i: jatorrizko datua; z_i: datu estandartua):

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

Aipatu behar da datu estandartuak dimentsiogabeak direla, hots, ez dute unitaterik.

4.4.1 Estandarketaren aplikazioa: datuak alderatzea

Datua estandartuak datu multzo ezberdinetako datuak alderatzeko erabil daitezke.

Adibidea: Ondoren, bi ikasle ezberdinek haien ikastetxeetan izandako Batxilergoko nota azaltzen da, ikastetxe horietan ikasleek azken urteotan izandako batezbesteko kalifikazioak eta kalifikazio horien desbideratze estandarrak:

Ikastetxeak	A	B
Ikasleen notak	7 (Mikel)	8 (Saioa)
Ikastetxeko batezbestekoak	6	8.5
Desbideratze estandarrak	0.5	1

Zein da erlatiboki kalifikazio altuena duen ikaslea?

Ikasleak ezin dira haien kalifikazioekin zuzenean alderatu, ikastetxe ezberdinetakoak direlako (baliteke B ikastetxea *eskuzabalagoa* izatea notak ematean). Behar bezala alderatzeko, bi ikasle horien notak estandartu behar dira:

$$z_{Mikel} = \frac{7-6}{0.5} = 2 \; ; \; z_{Saioa} = \frac{8-8.5}{1} = -0.5$$

Mikelek kalifikazio estandartu handiagoa du Saioak baino.

4.4.2 Estandarketaren aplikazioa: datu atipikoak

Batezbestekotik erlatiboki datuak zenbat aldentzen diren adierazten duenez, z balio estandartuak outlier edo datu atipikoak aurkitzeko ere erabil daiteke. Horretarako estatistikan eredu moduan maiz erabiltzen den banaketa normala (banaketa idealizatu bat, praktikan jasotzen diten datu multzoekin bat etortzen dena maiz) har daiteke oinarritzat.

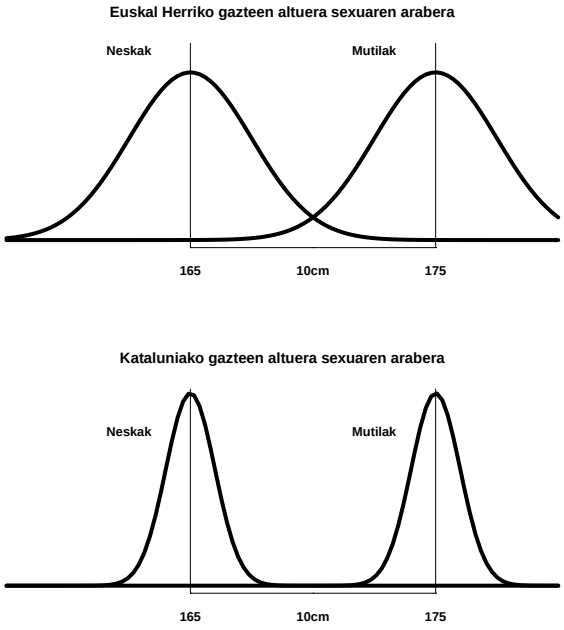
Datuak banaketa normalaren araberakoak direla suposatuz:

- ± 2.57 baino haratago dauden datuak ehun aldietatik behin bakarrik azaldu beharko lirateke;
- ± 3.09 baino haratago dauden datuak bostehun aldietatik behin bakarrik azaldu beharko lirateke;
- ± 3.29 baino haratago dauden datuak mila aldietatik behin bakarrik azaldu beharko lirateke;
- ± 3.72 baino haratago dauden datuak bost mila aldietatik behin bakarrik azaldu beharko lirateke;
- ± 3.89 baino haratago dauden datuak hamar mila aldietatik behin bakarrik azaldu beharko lirateke.

Datu atipikoetarako ezartzen dugun maiztasunaren arabera, datu atipikoak aurreko muga horietatik kanpo balio estandartuak dituzten datuak lirateke. Darabilgun datu multzoa baino lagin tamaina handiagoko mugak hartuko dira noski. Adibidez, 20 datu baditugu, datu atipikotzat har daitezke ± 2.57 daudenak (100etik behin gertatzen direnez), baina 3000 datu baditugu, gutxienez ± 3.72 kanpo daudenak hartu litezke atipikotzat.

4.5 Efektuaren tamaina

Bi datu-multzoetako batezbestekoak alderatzean, $\overline{x}_1 - \overline{x}_2$ aldea adierazgarria edo kontuan hartzekoa den ebaluatzea bilatzen da askotan. Alde jakin baterako, datuen bariantza zenbat eta txikiagoa orduan eta ziurtasun edo indar handiagoz baieztatu ahal izango da alde hori adierazgarria dela. Adierazgarritasun horri **efektuaren tamaina** deitzen zaio.



Irudia 4.2: Bi herrialdeetan sexuen arteko altuera aldea berdina bada ere, Katalunian efektuaren tamaina (aldearen garbitasuna) handiagoa da, sakabanatzea txikiagoa delako.

Efektuaren tamaina kalkulatzeko formula anitz dago, egoera eta suposizioen arabera. Hemen, bi datu-multzoek bariantza berdina dutela suposatuko dugu, lagin-errorea gorabehera; kasu horretan, hau da efektuaren tamaina aukeran dagoen neurrietako bat, *Cohen-en d* izenekoa:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

non s bi datu-multzoetako desbideratze estandar bateratua den:

$$s = \sqrt{\frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}}$$

$\hat{s}_1^2$ eta $\hat{s}_2^2$ bi datu-multzoetako lagin-bariantzak izanik, hurrenez hurren.

Cohen-en d 0.3 baino txikiagoa denean, batezbestekoen diferentzia ahula dela irizten da, 0.8ra bitartean ertaina eta hortik gora (1 baino handiagoa izan daiteke) sendoa.

4.6 Ariketak

1. Denda bateko salmentak jaso dira zenbait egunetan (milaka eurotan):

   12.3-14.6-18.7-15.4

   Kalkulatu desbideratze estandarra (zuzendu gabea eta zuzendua), formula luze nahiz laburrarekin. Emaitza interpretatu.

2. Maila bateko matematika kalifikazioak bildu dira:

Kalifikazioak	Ikasleak
2-3	5
3-4	14
4-5	22
5-6	26
6-7	18
7-8	12

   Populazio-bariantza eta kuartil arteko ibiltartea kalkulatu behar dira.

3. Lantegi bateko bi makinek eraldaketa-prozesu berdina burutzen dute. Makina bakoitzean pieza batzuk eraldatzeko denbora jaso da (segundutan):

   A makina: 45-48-50-42-54-58-45-42-52-48
   B makina: 36-40-42-62-48-54-66-34-40-56

   Datu atipikoen eragina baztertzen duen neurri estatistiko bat erabiliz, bi makinetako datuen sakabanatzea alderatu eta produktibitateari zein ekoizpen-plangintzari begira zein makina hobetsi behar den erabaki.

4. Oinarritzat banaketa normala hartuz, aurkitu datu 45 tamainako lagin honetan datu atipikoak, jakinda horien batezbestekoa 48.02 dela, eta desbideratzea 10.17:

   52.3 59.4 39.4 44.6 55.7 42.2 49.8 60.9 24.0 38.7 33.6 62.3 53.3 36.3 36.1
   51.9 42.6 51.3 47.7 50.8 54.8 52.0 42.8 43.3 30.6 71.2 53.3 48.2 45.5 47.8
   52.3 48.0 49.4 53.5 60.2 54.5 49.6 55.8 55.0 50.3 17.6 42.0 58.0 47.1 45.5

5. 670 datu jaso dira. Datu horien batezbesteko aritmetik sinplea 104 da, eta desbideratzea 10.6. Hiru handienak eta hiru datu txikienak ematen dira ondoren:

   - Handienak: 148-135-122
   - Txikienak: 69-72-75

   Datu atipikoak bilatu behar dira, banaketa normala eredutzat harturik.

6. Udako eta neguko hainbat egunetan denda bateko eguneko salmentak jaso dira (eurotan). Datuak taula honetan bildu dira:

Salmentak	Udako egunak	Neguko egunak
0-100	45	87
100-200	95	97
200-300	146	122
300-400	100	99
400-500	67	32

   (a) Aztertu zein sasoitan diren eguneko salmentak sakabanatuago datu guztiak erabiltzen dituen neurri bat erabiliz eta erabaki zein sasoitarako izango den salmenten aurresan bat fidagarriago.

   (b) Efektuaren tamaina ebaluatu, interpretatu eta eztabaidatu.