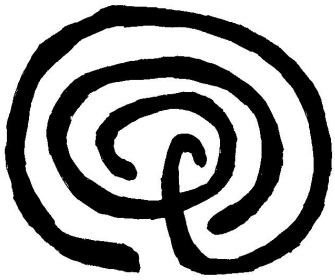
ESTATISTIKA ETA DATUEN ANALISIA

IV: Aldagai bakunaren deskribapena: sakabanatzea

Egilea: Josemari Sarasola



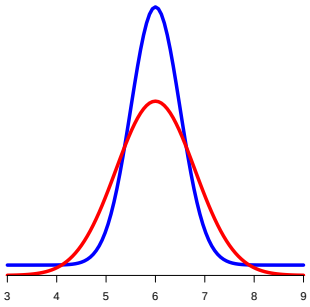
Gizapedia

gizapedia.hirusta.io

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.

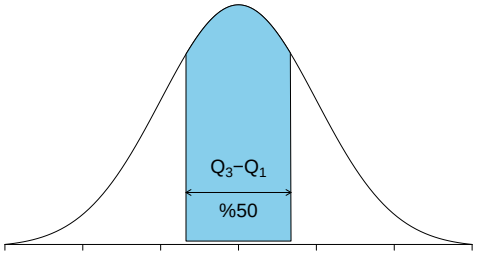
Baina hain zuzen ere sendoa ez izateagatik, datu atipikoak aurkitzeko tresna gisa erabiltzen da maiz, bereziki kalitate kontrolean.

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; datu guztiak berdinak direnean, hain zuzen), 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:

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

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.

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

22-25-28-26-24

Kalkulatu desbideratze estandarra eta bariantza 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

$$\overline{x} = \frac{125}{5} = 25min \ ; \ s_x^2 = \frac{3145}{5} - 25^2 = 4min^2; \ s_x = \sqrt{4} = 2min$$

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

4.2.3.2 Populazio-bariantza eta lagin-bariantza (eta desbideratze zuzendua)

Ikasi dugun bariantzaren aurreko formula **populazio-bariantzarena** da. Baina, estatistikan ohikoena laginak hartzea da, populazio osoaren ordez. Laginak jasota, horren emaitzak populazio osora zabaltzen dira, zenbatespen edo estimazio gisa. Populazio-bariantzaren emaitza populazio osora zabaltzean badago beti errore bat, lagin errorea deitzen duguna. Errore hori batez beste zuzentzeko, **lagin-bariantza** edo *bariantza zuzendua* erabiltzen da (ohartu s letrak txanoa daramala gainean):

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

Hortik, *desbideratze zuzendua* deitzen dena eratortzen da:

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

Populazio-bariantza erabiliko dugu datuak populazio osoko elementuei buruzkoak direnean edota besterik gabe lagin-errorea kontuan hartzea interesatzen ez zaigunean.

Ikusten denez, bien arteko diferentzia, betiere jatorriko formula harturik (bariantza zuzenduak ez du formula laburrik!), zati n edo zati $(n-1)$ egitea da. Lagin bariantza kalkulatzeko egin beharreko zuzenketa honi Bessel-en zuzenketa deritzo.

Edonola ere, bata ala bestea emanda, populazio- eta lagin-bariantzaren artean zuzeneko erlazio hau betetzen da, batetik bestera bihurtzeko aukera ematen duena:

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

Ikusten denez, lagin-bariantza beti da handiagoa populazio-bariantza baino. Lagin tamaina oso handia denean, bien arteko aldea oso txikia da (100 datuetarako esaterako, $n/(n-1) = 100/99 = 1.01$, %1ekoa), eta handixeagoa datu kopurua txoikia denean (4 datuetarako esaterako, $n/(n-1) = 4/3 = 1.33$, %33koa). Alde horiek logikoak dira: lagina txikia denean, lagin errorea handiagoa da, eta beraz zuzenketa handiagoa egin behar da bariantzaren emaitza populazio osora zabaltzean.

Estatistika-programa informatikoek bariantza eta desbideratze zuzenduak kalkulatzen dituzte gehienetan, besterik adierazten ez bada.

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

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

$$s_x^2 = \frac{15837.5}{22} - 26.36^2 = 25.03 urte^2 \rightarrow s_x = 5.004 urte$$

Beraz, batezbestez langile baten adina 5.03 urte desbideratzen da batez beste 26.36 urtekoko batezbestekotik.

Tartekako datuak ditugula, batezbestekoa eta bariantza kalkulatzeko klase-marka hartu dugu abiapuntu. Batezbestekoaren kasuan, klase-marka harturik ez da errore sistematikorik sortzen, datuak uniformekin banatuta tartean zehar, klase-marka hartzeagatik jatorrizko datuen ordez errore negatiboak errore positiboekin konpentsatzen direlako. Bariantzaren kasuan ordea, diferentzia guztiak ber bi egiten ditugunez, errore sistematiko bat sortzen da. Errore sistematiko hori neurri batean neutralitzeko Sheppard-en zuzenketa erabil daiteke, tarte-zabalera konstante denean:

$$s^2_{x,sheppard} = s^2_x - \frac{b^2}{12}$$

b tarte-zabalera izanik.

Ariketan, Sheppard-en zuzenketaz horrela doituko genuke bariantza:

$$s^2_{x,sheppard} = s^2_x - \frac{b^2}{12} = 25.03 - \frac{5^2}{12} = 22.95 \rightarrow s_x = 4.79 urte$$

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 Sakabanatzeak alderatuz: sakabanatze-neurri erlatiboak

Aurreko atalean ikasitako sakabanatze-neurri absolutuak ez dira egokiak datu-multzoak 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, sakabanatzea zentro-neurri edo kideko formula batekin alderatu behar dugu. Horrela **sakabanatze-neurri erlatiboak edo sakabanatze-koefizienteak** *ditugu (ingelesez,* **coefficient of dispersion***) ditugu:*

- **ibiltarte erlatiboa**: $RR = \frac{R}{\overline{x}}$
- **ibiltarteko sakabanatze-koefizientea**: $CD_R = \frac{R}{x_{min}+x_{max}} = \frac{x_{max}-x_{min}}{x_{min}+x_{max}}$
- **aldakuntza-koefizientea**: $A_X = \frac{s_x}{\overline{x}}$
- **kuartil arteko sakabanatze-koefizientea**: $CD_{IQR} = \frac{Q_3-Q_1}{Q_3+Q_1}$
- **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 datumultzo ezberdinen sakabanatzeak alderatzeko.

Zenbat eta handiagoak izan, sakabanatzea orduan eta handiagoa da, baina hala ere ezin da zehaztu sakabanatzeko handiko edo txikiko baliorik, sakabanatzea ezaugarri erlatiboa baita: beste datumultzo 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, neurri erlatiboen ordez.

4.4 Estandarketa eta puntuazio estandarrak

Estandarketa x_i datuak transformatu egiten dituen aldagai-aldaketa bat da, emaitza moduan z_i *puntuazio estandartuak* ematen dituena:

$$\underbrace{x_i}_{datuak} \longrightarrow \underbrace{z_i = \frac{x_i - \overline{x}}{s_x}}_{punt.\ estandarrak\ (z\ scores)}$$

Estandarketak dituen aplikazioen artean hauek ikasiko ditugu jarraian:

- aldagaiak dimentsiogabetu edo unitaterik gabe uztea,
- datu multzo desberdinetako datuak alderatzea,
- datu atipikoak atzematea.

4.4.1 Estandarketaren aplikazioak: aldagaiak dimentsiogabetzea

Estatistikan ohikoa da aldagai estatistikoak agregatu edo gehitu behar izatea, adibidez adierazleak eratzeko. Aurreko ikasgaian normalizazio izeneko prozesua ikasi genuen datuen agregazioa era egokian egiteko. Estandarketa da beste aukera bat aldagai eta magnitude desberdinetako datuak gehitu ahal izateko (gogoratu bestela sagarrak eta intxaurrak ezin direla besterik gabe gehitu).

4.4.2 Estandarketaren aplikazioak: datuak alderatu eta sailkatzea

Eskala berera bihurtutako datuak direla kontuan harturik, puntuazio estandartuak datu multzo ezberdinetako datuak alderatu eta sailkatzeko erabil daitezke.

Adibidea: Ondoren, bi ikasle ezberdinek haien ikastetxeetan izandako Batxilergoko nota azaltzen da, ikastetxe horietako ikasleen batezbesteko kalifikazioak eta 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:

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

Beraz, Mikelek kalifikazio estandar handiagoa du Saioak baino.

4.4.3 Estandarketaren aplikazioak: datu atipikoak

Datu atipikoek emaitza estatistikoetan eragiten duten distortsioa dela eta, garrantzitsua da horiek atzemateko irizpideak izatea. z balioak batezbestekotik erlatiboki datuak zenbat aldentzen diren adierazten duenez, puntuazio estandarrak *outlier* edo datu atipikoak aurkitzeko ere erabil daitezke, datu atipikotzat hartuz z balio absolutu jakin bat (gainetik nahiz azpitik, goiko nahiz beheko muturrean) gainditzen duten datuak.

Arazo bat sortzen da, ordea. z balioak kalkulatzeko, batezbestekoa eta desbideratze estandarra behar ditugu, eta haien kalkuluan datu atipikoek eragin nabarmena dute. Beraz, datu atipikoak aurkitzeko baliatzen ditugun puntuazio estandar arruntak kalkulatzerakoan, datu atipikoek nabarmen eragiten duten neurrietan oinarritzen gara. Kontraesan hori saihestu eta estandarketan datu atipikoek eraginik izan ez dezaten, z puntuazio estandar sendoak erabiltzen dira, zentroa medianaren bidez ematen duena duena; eta desbideratzearen ordez, $1.4826 \times DAME$ erabiltzen duena, betiere suposatuz banaketa normala, batezbestekotik alde banatara simetria duen banaketa estatistiko ideal bat, datuen eredu gisa erabil daitekeela:

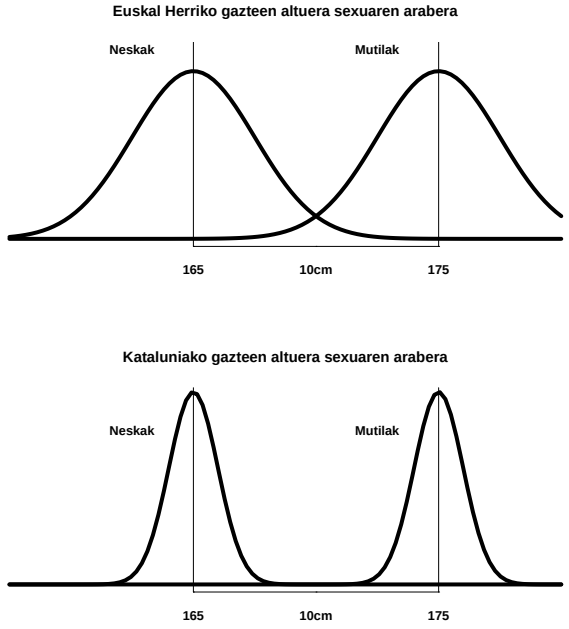
$$z_a = \frac{x_i - Me}{1.4826 \times DAME}$$

z puntuazio estandarrak nahiz sendoak erabilita, datu bat atipikoa dela baieztatzeko gainditu behar den muga-balioa datu atipiko bat agertzeko aldez aurretik ezarri nahi dugun probabilitatearen araberakoa da. Honako taula honetan, datu atipikoak agertzeko probabilitate desberdinetarako z muga-balio absolutuak zehazten dira, betiere banaketa normaletik harturik:

Prob(atipiko)	$\|z\|$
0.1	1.6449
0.05	1.96
0.02	2.3263
0.01	2.5758
0.005	2.807
0.001	3.2905

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 (*pooled standard deviation*) den:

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

s_1^2 eta s_2^2 bi datu-multzoetako lagin-bariantzak izanik, hurrenez hurren, eta n_1 eta n_2 horien tamainak.

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

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

   (b) Kalkulatu bariantza zuzendua (lagin-bariantza) Bessel-en zuzenketa aplikatuz zuzenean.

2. Kurtso bateko ikasleen matematika kalifikazioak bildu dira:

Kalifikazioak	Ikasleak
2-3	5
3-4	14
4-5	22
5-6	26
6-7	18
7-8	12
	97

   (a) Kuartil arteko ibiltartea kalkulatu behar da.
   (b) Kuartil arteko ibiltartea datuen banaketari dagokion histogramaren barnean irudikatu eta adierazi histograma zati horrek histogramako azalera osoaren zenbateko zatia hartzen duen portzentajean.
   (c) Populazio-bariantza eta populazio-desbideratzea kalkulatu eta interpretatu, Sheppard-en zuzenketa aplikatuz.

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

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

   Kuartil arteko ibiltartea nahiz DAME baliatuz
   (a) bi makinetako datuen sakabanatzea alderatu;
   (b) produktibitateari zein ekoizpen-plangintzari begira zein makina hobetsi behar den erabaki;
   (c) zein da bi neurri horiek partekatzen duten ezaugarria?

4. Bi ebaluatzaile ezberdinek lanpostu baterako hautagaiak (a, b, ..., i) baloratu dituzte ataza bat burutzeko erakutsi duten trebeziari buruz. Honako hauek dira datuak:

   - A ebaluatzailea: a:56-b:67-c:77-d:84
   - B ebaluatzailea: e:84-f:72-g:70-h:74-i:80

   Hautagai guztiak ordenatu behar dira haien puntuazioari buruz. (Soluzioa: A ebaluatzailea, $\sum x = 284, \overline{x} = 71, \sum x^2 = 20.610, s_x = 10.56$; B ebaluatzailea: $\overline{x} = 76, s_x = 5.22$)

5. 11 datu jaso dira:

$$82 - 60 - 72 - 75 - 85 - 91 - 93 - 101 - 122 - 135 - 148$$

(a) Datu atipikoak bilatu behar dira, puntuazio estandar sendoak kalkulatuz eta $z = 1.5$ muga-balioa harturik.
(b) z puntuazio muga berdina harturik baina oraingo honetan ohiko puntuazio estandarrak kalkulatuz, datu atipikoen zehaztapen berria egin ezazu, eta aurreko zehaztapenarekin duen aldeari buruzko azalpena egin ezazu. Laguntza: datuen batura 1064 da, eta datu karratuen batura 110.502.

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
	453	437

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

(b) Arestiko atalean kalkulatu dituzun estatistikoak harturik, bi banaketen baterako eskema bat egin ezazu, maiztasun kurbak marraztuz.

(c) Efektuaren tamaina ebaluatu, interpretatu eta eztabaidatu.

SAKABANATZEARI BURUZKO ARIKETEN EBAZPENAK

1. ariketa

x_i	x_i^2	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$

(a)

Lehen gauza, batezbestekoa: $\overline{x} = \dfrac{\sum x_i}{n} =$

s_x (zuzendu gabea)

FORMULA AZKARRA: $s_x = \sqrt{\dfrac{\sum x_i^2}{n} - \overline{x}^2} =$

JATORRIZKO FORMULA: $s_x = \sqrt{\dfrac{\sum (x_i - \overline{x})^2}{n}} =$

Beti positiboa da! Interpretazioa:

$\hat{s}_x$ (zuzendua)

• formula garatuz (zati (n-1) baina beti jatorriko formularekin!):

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

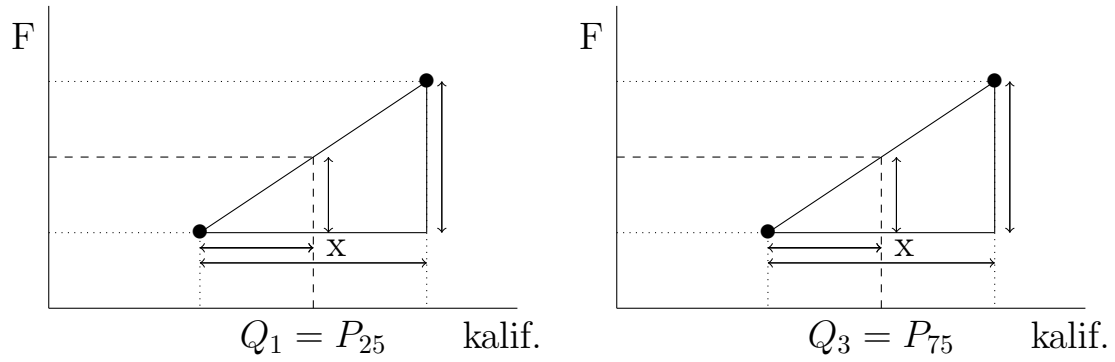
(b) Bessel-en zuzenketaren bitartez, hots,

$\hat{s}_x$ (zuzendua) populazio-bariantzatik (zuzendu gabeko bariantzatik)

$s_x^2 =$ $\qquad \rightarrow \hat{s}_x^2 = \dfrac{n}{n-1} s_x^2 =$ $\qquad \rightarrow \hat{s}_x = \sqrt{\qquad} =$

2. ariketa

(a)

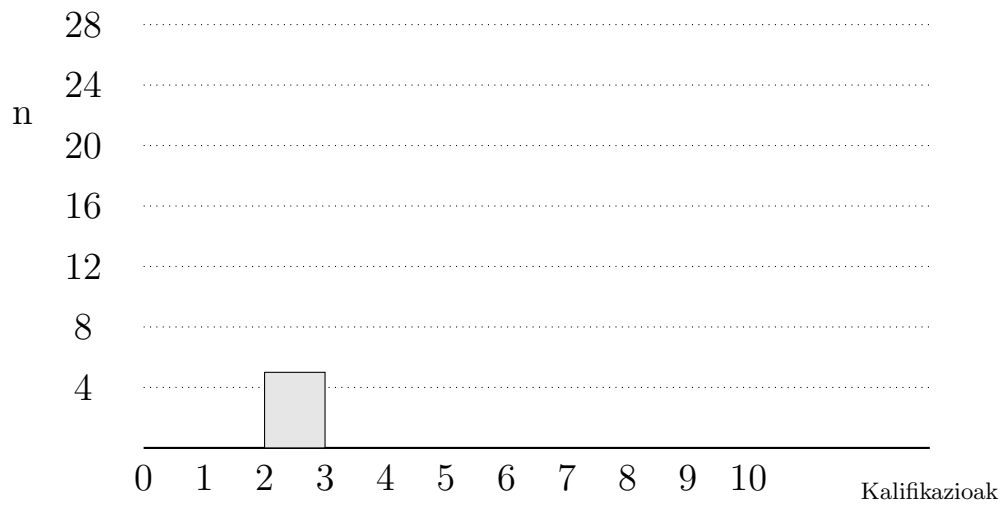


$$\frac{x}{\text{——}} = \text{——} \rightarrow x = \qquad \rightarrow Q_1 = P_{25} =$$

$$\frac{x}{\text{——}} = \text{——} \rightarrow x = \qquad \rightarrow Q_3 = P_{75} =$$

$$IQR = Q_3 - Q_1 =$$

(b)



(c) Formula laburra erabiliko dugu:

$$s_x^2 = \frac{\sum x^2}{n} - \overline{x}^2$$

x	n	nx	nx^2 $_{(nx)^2 \text{ ez!!}}$

Beti bezala, bariantza eta desbideratzea kalkulatzeko, lehen gauza batezbestekoa kalkulatzea (3. zutabean):

$\overline{x} =$ ———————— $=$

Bariantza: $s_x^2 =$ ———————— $-$ $=$

Sheppard-en zuzenketa aplikatuz:

$s^2_{sheppard,x} = s_x^2 - \frac{b^2}{12} =$

Desbideratzea: $s_x = \sqrt{\qquad\qquad\qquad} =$

3. ariketa

Lehenik eta behin kuartil arteko ibiltartea eta DAME kalkulatzeko mediana, lehenengo eta hirugarren kuartilak kalkulatu behar dira:

A makina

Me=erdiko datua=$\frac{\text{————————}}{2}$ =
$Q_1 = P_{25}?10 \times 0.25 = 2.5gn\ datua \rightarrow$

$Q_3 = P_{75}?10 \times 0.75 = 7.5gn\ datua \rightarrow$

Oharra: Aukeran, mediana $10 \times 0.5 = 5$gn datu gisa ere eman daiteke.

B makina

Me=erdiko datua=$\frac{\text{————————}}{2}$ =
$Q_1 = P_{25}?10 \times 0.25 = 2.5gn\ datua \rightarrow$

$Q_3 = P_{75}?10 \times 0.75 = 7.5gn\ datua \rightarrow$

$IQR(A) =$ $;IQR(B) =$

Kalkula dezagun DAME:

A makina

$DA = |x_i - Me| : 6 - 6 - .....$

DA_{ord} :

$DAME = Me(|x_i - Me|)?\ 10 \times 0.5 = 5gn\ DA_{ord} \rightarrow DAME_A =$

B makina

$DA = |x_i - Me|$:

DA_{ord} :

Bi makinetako sakabanatzeak alderatzeko **sakabanatze-neurri erlatiboa** baliatu behar da, arestian kalkulatutako neurriei dagokiena:

$\bullet IQR(A)/Q_3(A)+Q_1(A) =$ $; IQR(B)/Q_3(B)+Q_1(B) =$

Beraz, neurri honen arabera,

$\bullet DAME_A/Me_A =$ $; DAME_B/Me_B =$

Beraz, neurri honen arabera, aurrekoan bezala (ez du horrela zertan izanik)

(b)

Produktibitateari buruz, orokorrean denbora txikiena duen makina da hoberena. Horretarako, kalkulatuko dugun medianari erreparatuko diogu, eta horrela B makina da onena.

Ekoizpen plangintzari buruz, berriz, sakabanatze txikiena duena makina da hoberena, plangintzarako denbora aurreikuspen egonkor eta ziurrak egiteko. Alde horretatik, A makina da hoberena.

Bi irizpideak bateraturik, bi makinen artean dilema sortzen da eta beraz ezinda baieztatu besterik gabe zein den hoberena.

4. ariketa

gogoratu: $z = \dfrac{x - \overline{x}}{s_x}$

$x_i(A)$	$x_i^2(A)$	$z_i(A)$	$x_i(B)$	$x_i^2(B)$	$z_i(B)$
a:56			e:84		
b:67			f:72		
c:77			g:70		
d:84			h:74		
-			i:80		
284					

$$\overline{x}_A = \frac{284}{4} = 71 \qquad ; \quad \overline{x}_B = \frac{\qquad\qquad}{5} =$$

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

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

Hautagaien ordenazioa:

1	2	3	4	5	6	7	8	9

5. ariketa

(a)

$x(ord)$	60										
$\lvert x - Me \rvert$											
$\lvert x - Me \rvert (ord)$											
$z_a = \frac{x - Me}{1.4826 DAME}$											

$Me =$; $DAME = 16$

(b)

$\overline{x} = \frac{\qquad}{11} =$

$s_x = \sqrt{\frac{\sum x_i^2}{n} - \overline{x}^2} = \sqrt{\frac{\qquad}{11} - \qquad} =$

x_i	82	60	72	75	85	91	93	101	122	135	148	$\sum =$
x_i^2												$\sum =$
$z_i = \frac{x_i - \overline{x}}{s_x}$												$\sum =$ $0(beti)$

Iruzkina: Puntuazio estandar arruntek eta puntuazio estandar sendoek, datuak atipikoak ezartzeko muga-balio berdina harturik ere, balio atipiko ezberdinak zehazten dituzte. Logikoa da, metodo desberdinak direlako biak. Egokiena puntuazio estandar sendoena dela esango genuke, kontraesankorra baita berez datu atipikoak aurkitzeko datu atipikoen mendean dauden puntuazio estandar arruntak baliatzea.

6. ariketa

(a)

x	n_{uda}	$n_{uda}x$	$n_{uda}x^2$	n_{negu}	$n_{negu}x$	$n_{negu}x^2$
	45			87		
	95			97		
	146			122		
	100			99		
	67			32		

Datu guztiak kontuan hartzen dituen neurria desbideratze estandarra da:

$$\overline{x}_{uda} = \text{————} = \qquad ; \quad \overline{x}_{negu} = \text{————} =$$

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

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

Sakabanatzeak alderatzeko neurri erlatiboa behar dugu; kasu honetan, desbideratzeari dagokiona aldakuntza-koefizientea da:

$$A_{uda} = \text{————} = \qquad ; \quad A_{negu} = \text{————} =$$

Beraz, salmentak sakabanatuagoak dira .....

Salmenten aurresanak fidagarriagoak izango dira ....., sakabanatze txikiagoa izanik, sasoi horretan salmentak egonkorragoak baitira.

(b)

---

salmentak

(c) Efektuaren tamaina neurtzeko Cohen-en d kalkulatzen dugu, suposatuz betiere, lagin bariantzak desberdinak diren arren, populazio mailan berdinak direla.

Lehenik eta behin desbideratze partekatu edo bateratua (*pooled deviation*) kalkulatzen dugu:

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\text{————————————}} = 120.2$$

Eta azkenik Cohen-en d kalkulatzen dugu:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s} = \text{————————————} = 0.29$$

Interpretazioa: Bi sasoietako eguneko salmenten batezbestekoen arteko diferentzia ez da garbia, ahula baizik.