

# Estatistika Deskribatzailea

2004/2005 ikasturtea

## 1. Aldagai Estatistikoak

Populazio edo lagin batean ezaugarri bat edo aldagai estatistiko bat behatu edo neurtzen denean datu edo balio estatistikoak lortzen dira, aldagai estatistiko horrek populazio edo lagin horretan hartzen dituen balio estatistikoak, hain zuzen ere.

Aldagai estatistikoak horrela sailkatzen ohi dira:

*aldagai kualitatiboak:* Aldagai hauek, populazio edo lagin batean, kualitatezko ezaugarriak neurtzen dituzte. Esate baterako, pertsona multzo batean, sexua edo ilearen kolorea neurtzen duten aldagaiak. Aldagia hauek hartzen dituzten balioak ezin dira ordenatu aritmetikoki eta, beraiekin, ezin dira eragiketarik egin.

*aldagai ordinalak:* Aldagia hauek hartzen dituen balioak ordena daitezke baina, beraien arteko diferentzia finkatuta egon gabe. Inkesta batzuetan, planteaturiko item-ekiko adostasuna, 0 eta 5 bitarteko zenbaki batez adierazi behar da; esate baterako, *Gobernuak langabeziaren problema konpontzeko, ahal duen guztia egiten du* esaldiarekin adostasun osoa adierazteko, 5eko bat jarriko dugu, 4ko bat, adostasun handia adierazteko, ..., 0 bat ezadostasun osoa adierazteko. Baina balio hauen arteko diferentziak ez dira berdinak, ezin dira aritmetikoki kontsideratu.

*aldagai kuantitatiboak:* Aldagia hauek hartzen dituzten balioak zenbakarriak edo neurgarriak dira. Esate baterako, pertsona multzo batean, seme-alaben kopurua, hilabeteen harturiko aspirina kopurua, edadea, pisua, altuera edo kolesterola neurtzen dituzten aldagaiak. Aldagia hauetako batzuk *diskretuak* direla esaten da, hartzen dituen balioak zenbaki isolatuak direlako, bitarteko balioak hartu ezinik: pertsona batek bi edo hiru seme-alaba izan dezake baina 2.5 ez; beste pertsona batek bi, hiru edo agian, bi aspirina eta erdia har dezake baina, 2.37 aspirina ez. Beste aldagai kuantitatibo batzuk, berriz, jarraituak direla esaten da, hartzen dituzten balioak tarte bateko edozein zenbaki erreal izan daitezkeelako. Pertsona baten pisua, esate baterako, 70 edo 72 kg-koa izan daiteke baita, 71.846 kg-koa ere.

## 2. Maiztasunak eta Adierazpide Grafikoak

Problema estatistiko baten datuek, bata bestearen atzetik ikusita, izugarritzko itxura izan dezakete, edonoren adorea kentzeko modukoa.

### Adibidea 1: Problema estatistiko baten datuak

*Ondoko taulan, 10 urte baino gehiago elkarrekin daramaten 150 bikoteren seme-alaben kopuruak agertzen dira*

0	0	1	1	2	0	3	0	2	4	5	3	2	2	1
2	0	6	3	2	1	1	1	0	0	5	1	1	2	3
1	0	4	5	2	1	2	1	1	2	2	2	2	4	1
2	3	4	5	1	3	4	5	2	1	4	1	2	1	2
1	1	1	0	0	2	0	2	0	2	1	3	0	4	1
1	1	0	1	0	2	3	3	1	4	4	1	6	1	0
3	0	0	0	1	1	1	2	3	3	3	1	2	3	0
6	5	1	2	0	0	3	6	2	1	1	1	4	1	2
4	5	1	2	1	2	2	4	2	3	3	3	4	3	1
5	5	2	3	1	3	1	5	0	0	0	4	2	0	5

Horrexegetatik, Estatistikak metodo batzuk bilatzen ditu datu horiek murrizteko edo, behintzat, laburtzeko eta dogokien esangura era errazagoan eman ahal izateko.

## 2.1. Maiztasunak

*Maiztasun absolutua:* Balio baten maiztasun absolutua, balio hori agertzen den aldi kopurua da ( $x$  balioaren maiztasun absolutua,  $F_x$  ikurraz adierazten ohi da) Hala ere, maiztasun absolutuak ez dira baliogarriak tamainu desberdineko bi populazio alderatu ahal izateko. Komenigarria da, beraz, populazioaren tamainuarekiko balio erlatiboak lortzea.

*maiztasun erlatiboa:* Maiztasun absolutuak, populazioaren tamainuaz zatitzen badira, balioen maiztasun erlatiboak lortzen dira (populazioaren tamainua  $n$  bada,  $x$  balioaren maiztasun erlatiboa,  $f_x = \frac{F_x}{n}$  izango da)

*maiztasun metatuak:* Batzutan,  $x$  balio bat emanda, zenbat balio  $\leq x$  lortu diren edo  $\leq x$  balioak zein ehunekotan ematen diren jakitea interesatzen zaigu. Lehenengo kasuan, erantzuna da  $x$  balioaren maiztasun absolutu metatua, hau da,  $\sum_{y \leq x} F_y$  eta bigarrean,  $x$  balioaren maiztasun erlatibo metatua, hau da,

$$\sum_{y \leq x} f_y$$

### Adibidea 2: Datuen antolaketa

Aurreko adibidean balio bakoitza zenbat aldiz agertzen den kontuan harturik, ondoko **maiztasun taula** eraiki daiteke:

Balioa	Maiztasun absolutua	M. absolutu metatua	Maiztasun erlatiboa	M. erlatibo metatua
0	26	26	0.173	0.173
1	42	68	0.280	0.453
2	32	100	0.213	0.667
3	21	121	0.140	0.807
4	14	135	0.093	0.900
5	11	146	0.073	0.973
6	4	150	0.027	1.000
<i>Guztira</i>	150		1.00	

## 2.2. Adierazpide grafikoak

Askotan, maiztasun taulen adierazpide grafikoak lagungarriak dira datuen propietateak hobeto ulertzeko. Estatistikako liburuetan grafiko mota asko ikus daitezke baina erabilgarrienak “Sektore-Diagramak ”eta “Barra-Diagramak ”dira (ikusi lehen adibideari dagozkion adierazpen grafikoak)

## 2.3. Aldagia jarraituak

Orain arte ikusitako adibideetan agertzen ziren aldagai estatistikoak diskretuak izan direnez, ikus dezagun aldagai estatistiko jarraitu bati dagokion adibidea

### Egoera 1: Beste problema estatistiko baten datuak

*Landare baten 250 hazi, gramotan pisatu ondoren, datu hauek lortzen dira:*

4.18	4.50	3.76	3.33	2.61	3.25	5.35	2.11	4.47	6.56	2.80	2.44
5.71	4.96	4.64	4.34	7.06	4.77	3.43	2.90	3.42	3.93	3.67	3.52
3.24	4.99	5.27	4.12	5.41	5.22	5.06	4.40	4.61	5.75	4.76	3.28
3.51	4.41	5.32	5.24	3.80	5.09	3.71	5.37	6.04	4.81	4.18	3.37
2.77	4.24	4.49	4.32	4.77	4.06	3.38	1.70	3.30	5.92	2.87	5.61
3.86	3.39	4.46	4.83	4.34	3.21	2.68	4.87	5.70	3.94	3.85	3.16
3.96	4.88	4.52	3.14	3.03	3.29	2.89	4.24	3.66	3.36	3.91	3.72
4.07	5.51	3.04	3.82	3.02	4.77	3.93	3.76	4.02	5.05	3.18	2.98
3.05	3.42	5.31	4.69	3.86	3.80	4.15	4.29	3.87	4.21	5.15	1.74
1.78	6.66	4.23	4.92	4.01	4.27	2.93	2.94	3.54	1.90	3.37	3.27
2.23	4.49	1.93	3.24	4.78	3.44	4.93	3.40	5.45	4.47	4.81	3.50
4.22	4.41	3.28	3.96	3.64	3.97	4.76	4.11	4.51	6.97	3.19	4.35
4.05	3.47	3.02	3.76	3.38	4.29	3.68	2.80	3.21	3.70	4.30	4.65
3.42	4.69	4.20	4.31	3.75	3.74	3.57	2.35	5.50	3.79	4.48	3.58
2.36	4.42	2.65	4.66	5.38	3.09	3.27	3.45	4.54	4.68	4.46	4.18
4.35	3.95	3.02	3.78	3.50	4.51	2.16	4.97	3.41	4.26	4.08	4.13
2.67	4.70	3.06	4.12	5.48	4.55	4.56	3.97	5.34	3.23	5.52	5.83
1.88	3.09	6.72	5.28	4.43	3.91	2.51	3.96	4.47	3.53	3.84	2.99
2.55	5.35	4.41	3.67	2.40	5.65	3.89	4.33	4.06	4.20	2.81	3.80
6.00	3.03	4.42	3.55	5.51	3.44	3.92	2.01	4.38	3.24	4.76	3.01
4.01	4.71	4.22	4.95	3.03	4.68	1.51	2.87	5.29	3.84		

Balioen maiztasuna aztertu baino lehen, balioak berak bildu edo elkartzea komenigarria izango da

### Egoera 2: Datuen antolaketa

*Datuak hainbat klasetan sailkatu ondoren, aurreko kasuaren antzera, taula hau eraiki daiteke:*

<i>klaseak</i>	<i>Erdiko puntuak</i>	<i>Maiztasun absolutua</i>	<i>Maiztasun erlatiboa</i>
(1.5,2.5]	2	15	0.06
(2.5,3.5]	3	64	0.256
(3.5,4.5]	4	99	0.396
(4.5,5.5]	5	54	0.216
(5.5,6.5]	6	13	0.052
(6.5,7.5]	7	5	0.02
		250	1.000

### 3. Aldagaien Zenbakizko Deskribapena

Populazio edo lagin batean behaturiko aldagai estatistiko baten datuak, zenbaki esangarri bakar batzuen bidez laburtzea beharrezkoa da. Zenbaki hauei, estatistikoak deitzen zaie eta, zein balioen inguruan datuak biltzen diren neurtzen dute edo datuen sakabanapena.

#### 3.1. Joera zentralerako estatistikoak

Demagun  $n$  tamainuko populazio edo lagin batean,  $X$  aldagai estatistiko batek  $x_1, x_2, \dots, x_n$  balioak hartzen dituela. Orduan,  $X$  **aldagaiaren batezbestekoa** horrela definitzen da,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Hala ere,  $n$  balio horietatik, bakarrik  $m$  desberdinak badira,  $x_1, x_2, \dots, x_m$ , batezbestekoa horrela idatz daiteke,

$$\bar{x} = \frac{\sum_{i=1}^m x_i F_i}{n} = \sum_{i=1}^m x_i \frac{F_i}{n} = \sum_{i=1}^m x_i f_i$$

non,  $F_i$  eta  $f_i$ ,  $x_i$  balioaren maiztasun absolutua eta erlatiboak diren

*Moda*, *mediana* eta *pertzentilak*, liburuetan agertzen diren joera zentralerako beste estatistiko batzuk dira.

#### 3.2. Datuen sakabanapena neurtzen duten estatistikoak

Argi dago, batezbestekoak, aldagiak hartzen dituen balio guztiak errepresentatu edo ordezkatu egiten dituela eta, ordezkapen hori, balioak kontzentratu badaude, ona izango dela eta, oso hedatuta badaude, ez dela hain ona izango. Esate baterako, ez da gauza bera batezbesteko nota 6koa izatea, 5, 6 eta 7ko notak atera ondoren edo 4, 4 eta 10eko notak atera ondoren.

Batezbestekoarekiko, datuen desbidazioa kalkulatzeko baditugu,

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$$

baina batezbesteko desbidazioa,

$$\frac{\sum_{i=1}^n (x_i - \bar{x})}{n} = 0$$

izango da eta hau baliogarria ez denez, desbidazioen karratuen batezbestekoa kalkulatu da,  $(\sigma^2)$ , **bariantza** deitzen dena,

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2$$

Balio desberdinak eta dagozkien maiztasunak, bakarrik, kontuan hartzen baditugu, bariantzarako formulak horrela geratuko dira,

$$\sigma^2 = \sum_{i=1}^m (x_i - \bar{x})^2 f_i = \sum_{i=1}^m x_i^2 f_i - \bar{x}^2$$

Batzutan, datuen sakabanapena neurtzeko **desbidazio tipikoa** izeneko estatistikoa ( $\sigma \geq 0$ ) erabiltzen da, hau da, bariantzaren erro karratua dena. Hala ere, eta neurketa unitateak sahiesteko, **aldakuntza koefizientea** erabiltzen da, hau da,  $\frac{\sigma}{|\bar{x}|}$  koefizientea edo ehunekotan adierazita,  $\frac{\sigma}{|\bar{x}|} \times 100$

## 4. Aldagai bidimentsionalak

Askotan, populazio edo lagin batean, ezaugarri bat baino gehiago aztertzea interesatzen zaigu. Kasu hauetan, hainbat dimentsiotako aldagai estatistikoez hitz egingo dugu. Kontzeptuak finkatzeko, suposa dezagun  $n$  tamainuko populazio edo lagin batean, bi aldagai,  $X$  eta  $Y$ , aztertu nahi direla, hau da,  $(X, Y)$  aldagai bidimentsionala aztertu nahi dela.

$X$  aldagaiak hartzen dituen balio desberdinak,  $x_1, x_2, \dots, x_{n_1}$  badira eta  $Y$  aldagaiarenak  $y_1, y_2, \dots, y_{n_2}$ ,  $(X, Y)$  aldagai bidimentsionalaren balio bat  $(x_i, y_k)$  horrelako bikote bat izango da eta, orain arte bezala, balioaren maiztasun absolutua ( $F_{ik}$ ) eta maiztasun erlatiboa ( $f_{ik} = F_{ik}/n$ ) kalkulatu daitezke.

$X$  aldagaiaren  $x_i$  balioaren maiztasun absolutua eta erlatiboa  $F_i^X$  eta  $f_i^X$  ikurrez adierazten badira,  $1 \leq i \leq n_1$ ; eta  $Y$  aldagaiaren  $y_k$  balioarenak,  $F_k^Y$  eta  $f_k^Y$  ikurrez,  $1 \leq k \leq n_2$ , ondokoa betetzen da (ikus adibidea):

$$F_{1k} + F_{2k} + \dots + F_{n_1 k} = \sum_{i=1}^{n_1} F_{ik} = F_k^Y$$

$$F_{i1} + F_{i2} + \dots + F_{i n_2} = \sum_{k=1}^{n_2} F_{ik} = F_i^X$$

eta gauza bera maiztasun erlatiboekin,

$$\sum_{i=1}^{n_1} f_{ik} = f_k^Y \quad \text{eta} \quad \sum_{k=1}^{n_2} f_{ik} = f_i^X$$

Batzutan, aldagi baten balio konkretu baterako beste aldagia nola banatzen den jakitea interesatzen zaigu.  $Y|_{X=x_i}$  ikurraz,  $X$  aldagaiaren  $x_i$  baliorako,  $Y$  aldagaia adierazten da eta  $X = x_i$  baldintzapean  $Y$  aldagaiaren  $y_k$  balioaren maiztasun absolutua, hau da,  $Y = y_k|_{X=x_i}$  egoeraren maiztasun absolutua,  $(x_i, y_k)$  balioarena izango da, hain zuzen,  $F_{ik}$ .  $Y = y_k|_{X=x_i}$  balioaren maiztasun erlatiboa, berriz,  $\frac{F_{ik}}{F_i^X} = \frac{f_{ik}}{f_i^X}$

### Adibidea 3: Aldagai bidimentsional baten adibidea

**taula 1** *Taula honetan, herri bateko biztanleen altuera eta pisua aldagaien banaketa bidimentsionala ematen da, maiztasun erlatiboen bidez.*

altuera → pisua ↓	1.55	1.60	1.65	1.70	1.75	1.80	1.85	1.90	1.95	2.00
50	.0007	.0037	.0099	.0144	.0114	.0051	.0013	.0002	.0000	.0000
55	.0007	.0040	.0122	.0198	.0175	.0086	.0025	.0004	.0000	.0000
60	.0006	.0039	.0136	.0248	.0244	.0132	.0041	.0007	.0001	.0000
65	.0004	.0034	.0137	.0282	.0309	.0185	.0062	.0012	.0001	.0000
70	.0003	.0027	.0125	.0291	.0356	.0236	.0087	.0018	.0002	.0000
75	.0002	.0020	.0103	.0272	.0373	.0274	.0111	.0025	.0003	.0000
80	.0001	.0013	.0077	.0230	.0356	.0291	.0129	.0032	.0005	.0000
85	.0000	.0007	.0052	.0177	.0309	.0281	.0138	.0037	.0006	.0000
90	.0000	.0004	.0032	.0124	.0244	.0248	.0134	.0040	.0007	.0001
95	.0000	.0002	.0017	.0078	.0135	.0200	.0120	.0039	.0007	.0001
100	.0000	.0001	.0009	.0045	.0114	.0147	.0098	.0035	.0007	.0001
105	.0000	.0000	.0004	.0023	.0068	.0098	.0075	.0029	.0006	.0001
110	.0000	.0000	.0002	.0011	.0037	.0060	.0050	.0021	.0005	.0001
115	.0000	.0000	.0001	.0005	.0018	.0033	.0031	.0015	.0004	.0000

**taula 2** *Bigarren taula honetan, pisuaren banaketa marginala erakusten da.*

<i>pisua</i>	50	55	60	65	70	75	80
<i>m. erlatiboa</i>	.0467	.0657	.0854	.1026	.1145	.1183	.1134
<i>pisua</i>	85	90	95	100	105	110	115
<i>m. erlatiboa</i>	.1007	.0834	.0599	.0457	.0304	.0187	.0107

**taula 3** *Beste honetan berriz, altueraren banaketa marginala.*

<i>altuera</i>	1.55	1.60	1.65	1.70	1.75
<i>m. erlatiboa</i>	.0030	.0224	.0916	.2128	.2852
<i>altuera</i>	1.80	1.85	1.90	1.95	2.00
<i>m. erlatiboa</i>	.2322	.1114	.0316	.0054	.0005

**taula 4** Azken honetan, 1.85 altuerarako, pisuaren baldintzazko banaketa

<i>pisua</i>	50	55	60	65	70	75	80
<i>m. erlatiboa</i>	.0117	.0224	.0368	.0557	.0781	.0996	.1158
<i>pisua</i>	85	90	95	100	105	110	115
<i>m. erlatiboa</i>	.1239	.1203	.1077	.0880	.0673	.0449	.0278

## 5. Aldagaien arteko korrelazioa

Populazio edo lagin batean definituriko aldagai estatistikoak, askotan, erlazionatuta daude eta Estatistikaren helburu bat, aldagaien arteko erlazioak aurkitzea da. Erlazio hauek, *deterministak* izan daitezke, hau da, lege baten arabera (esate baterako, gorputzen erorketak aztertzen direnean, altuera ezagututa, lege batek, lurrera erortzean gorputzak daraman abiadura ematen digu). Baliteke, erlaziorik ez egotea eta orduan, aldagiak *independentek* edo *elkarrekiko askeak* direla esaten da (begien kolorea eta oinaren tamainua, adibidez) Gehienetan, nahiz eta ez ezagutu, erlazioen bat dagoela susmatzen da (aurreko adibideko pisu eta altueraren kasuan bezala)

$X$  eta  $Y$  aldagiak erlazio edo korrelazio positiboa daukatela esaten da, populazioaren elementu batean  $X$  aldagiak balio handienetarikoa bat hartzen duenean,  $Y$  aldagiak ere, balio handienetarikoa bat hartzen bada eta alderantziz,  $X$  aldagiak balio txikienetarikoa bat hartzen duenean,  $Y$  aldagiak ere bai. Guztiz kontrakoa gertatzen bada, hau da,  $X$  aldagiak balio handienetarikoa bat hartzen duenean,  $Y$  aldagiak balio txikienetarikoa bat hartzen bada, erlazioa edo korrelazioa negatiboa dela esaten da.

### 5.1. Kobariantza eta Pearson-en koefizientea

Aldera dezagun  $X$  aldagiak hartzen duen balio bakoitza, populazioaren elementu berean  $Y$  aldagiak hartzen duenarekin.

<i>Elementuak</i>	1	2	3	...	n
<i>X aldagaia</i>	$x^1$	$x^2$	$x^3$	...	$x^n$
<i>Y aldagaia</i>	$y^1$	$y^2$	$y^3$	...	$y^n$

Balioen desbidazioak, ondoko hauek izango dira,

$$(x^1 - \bar{x}), (x^2 - \bar{x}), \dots, (x^n - \bar{x}) \text{ eta}$$

$$(y^1 - \bar{y}), (y^2 - \bar{y}), \dots, (y^n - \bar{y}),$$

$X$  eta  $Y$ -ren arteko korrelazio positiboa egonda,  $(x^i - \bar{x}) > 0$  bada,  $(y^i - \bar{y})$  ere, gehienetan, positiboa izango da eta  $(x^i - \bar{x}) < 0$  bada,  $(y^i - \bar{y})$  ere, gehienetan, negatiboa, hau da,  $(x^i - \bar{x})$  eta  $(y^i - \bar{y})$  zeinu berbera izango dutela gehienetan. Korrelazioa negatiboa izanda berriz, guztiz alderantzizkoa, hau da,  $(x^i - \bar{x})$  eta  $(y^i - \bar{y})$  kontrako zeinua izango dute gehienetan.

Ondoko estatistikoak, desbidazioen arteko biderkaduren baturak,

$$(x^1 - \bar{x})(y^1 - \bar{y}) + (x^2 - \bar{x})(y^2 - \bar{y}) + \dots + (x^n - \bar{x})(y^n - \bar{y})$$

$X$  eta  $Y$  aldagaien artean zer nolako korrelazio dagoen adieraziko digu. Populazioaren tamainuaren eragina sahiesteko, aurreko batura aipaturiko tamainuaz zatitzen da eta lortzen den estatistikoari,  $X$  eta  $Y$  aldagaien arteko **kobariantza**,  $\sigma_{XY}$ , deitzen zaio. Beraz,

$$\begin{aligned}\sigma_{XY} &= \frac{\sum_{i=1}^n (x^i - \bar{x})(y^i - \bar{y})}{n} \\ &= \frac{\sum_{i=1}^n x^i y^i}{n} - \bar{x} \frac{\sum_{i=1}^n y^i}{n} - \bar{y} \frac{\sum_{i=1}^n x^i}{n} + n \frac{\bar{x} \bar{y}}{n} \\ &= \frac{\sum_{i=1}^n x^i y^i}{n} - \bar{x} \bar{y}\end{aligned}$$

Adibidean bezala, datuak,  $(x_i, y_k)$  bikoteen maiztasun erlatiboak ( $f_{ik}$ ) izango balira,

$$\sigma_{XY} = \sum_{i=1}^{n_1} \sum_{k=1}^{n_2} x_i y_k f_{ik} - \bar{x} \bar{y}$$

Bestalde, **Pearson-en korrelazio koefizientea**,  $\rho_{XY}$  ikurraz adierazita, horrela definitzen da,

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Edozein kasutan, konproba daiteke,  $-1 \leq \rho_{XY} \leq 1$  dela eta horrez gain,

- a)  $|\rho_{XY}| = 1$  bada, korrelazioa determinista dela
- b)  $\rho_{XY} > 0$  bada, korrelazio positiboa dagoela eta  $\rho_{XY} < 0$  bada, korrelazio negatiboa.
- c)  $|\rho_{XY}| \geq 0,8$  bada, korrelazio handia dagoela
- d)  $0,5 \leq |\rho_{XY}| \leq 0,8$  bada, korrelazio onargarria eta
- e)  $|\rho_{XY}| < 0,5$  bada, korrelazioa ahul samarra dela

## 6. Erregresio Lineal Sinplea

Erregresio lineala,  $X_1, X_2, \dots, X_m$  aldagaietatik abiatuak, beste  $Y$  aldagai bat estimatzeko edo iragartzeko metodo estatistiko bat da. Iragarri nahi den  $Y$  aldagaiari, menpeko aldagaia deitzen zaio eta  $X_i$  aldagaiei,  $Y$ -rako iragarpena egiteko erabiltzen direnei, *iragarleak* deitzen zaie. Aldagai iragarle bakarra dagoenean, erregresio lineal sinpleaz hitz egiten da, bestela, erregresio anizkoitzaz. Bestalde, erregresio linealaren ezaugarri nagusia da,  $Y$  aldagaiaren iragarpena funtzio lineal baten bidez ematen dela, hau da,  $a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b$  funtzio baten bidez erregresio anizkoitzaren kasuan eta,  $aX + b$  funtzio baten bidez erregresio sinplearen kasuan.

$X$  aldagaiak populazioaren edo laginaren  $i$ . elementuan  $x_i$  balioa hartzen badu,  $\tilde{y}_i = ax_i + b$ ,  $i$ . elementurako  $Y$  aldagaiaren balio estimatua edo iragarria izango da ( $Y$  aldagaiak  $i$ . elementuan hartzen duen benetako balioa  $y_i$  ikurraz adieraziko dugu)



## 6.1. Problemaren planteamendua

Populazio edo lagin batean,  $X$  eta  $Y$  aldagaierekiko,  $n$  behaketa egin ondoren, demagun  $(x_i, y_i)$ ;  $1 \leq i \leq n$ , balioak lortzen direla.  $\tilde{Y} = aX + b$ , bezalako zuzenaren ekuazioa aurkitu nahi da non,  $\tilde{y}_i = a x_i + b$ ,  $i$ . balio estimatua bada,  $(y_i - \tilde{y}_i)$ ;  $1 \leq i \leq n$ , ahalik eta txikienak diren, hau da,  $(x_i, y_i)$  puntuei hoberen doitzen zaien zuzena.

$\tilde{Y} = aX + b$  zuzenari  $X$  **gaineko**,  $Y$  **-ren erregresio zuzena** deitzen zaio. Erregresio zuzena aurkitzeko, hau da,  $a$  eta  $b$  parametroak aurkitzeko, *minimo karratuen metodoa* erabiliko da.

## 6.2. Minimo karratuen metodoa

$i$ . balio iragarria  $\tilde{y}_i = a x_i + b$  bada, puntu honetako iragarpenaren errorea,  $d_i = y_i - \tilde{y}_i = y_i - a x_i - b$  izando da.  $d_i$  erroreak ahalik eta txikienak izan daitezten,

a) ideia posible bat  $\sum_{i=1}^n d_i$  minimizatzea izan daiteke.

**Adibidea 4 :**

$(1, 1), (2, 1), (3, 1)$  balioak emanda,  $\tilde{Y}_1 = X - 1$  eta  $\tilde{Y}_2 = 3 - X$  zuzenek,  $\sum_{i=1}^3 d_i = 0$  betetzen dute baina, argi dago infinitu zuzen daudela, aurrekoek baino doikuntza hobeak sortzen dituztenak.

b) beste aukera  $\sum_{i=1}^n |d_i|$  minimizatzea izan daiteke.

**Adibidea 5 :**

$(2, 3), (4, 8), (6, 7)$  balioak emanda,  $\tilde{Y}_1 = X + 2$  zuzena hartuta,  $\sum_{i=1}^n |d_i| = 4$  eta  $\tilde{Y}_2 = 3 - X$  zuzenarekin,  $\sum_{i=1}^n |d_i| = 3$  Nahiz eta errorea handiagoa izan, argi dago lehen zuzenarekin lortzen den doikuntza, bigarrenarekin lortzen dena baino hobe delat.

c) soluzio ongarria  $E = \sum_{i=1}^n d_i^2$ , hau da, errore koadratikoen batura minimizatzea da.

$$E = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a x_i - b)^2$$

Gogora dezagun  $a$  eta  $b$  parametroak aurkitu nahi dira,  $E = \sum_{i=1}^n (y_i - a x_i - b)^2$  ahalik eta txikiena izan dadin eta horretarako,  $a$  eta  $b$  aldagaiekiko,  $E$  espresio edo funtzioaren minimoa aurkitu beharko da, hau da, ondokoa bete behar da,

$$\frac{\partial E}{\partial a} = 0 \quad \text{eta} \quad \frac{\partial E}{\partial b} = 0$$

$$\frac{\partial E}{\partial a} = 0 \iff \sum_{i=1}^n -2x_i(y_i - a x_i - b) = 0 \iff$$

$$\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i$$

eta,

$$\frac{\partial E}{\partial b} = 0 \iff \sum_{i=1}^n -2(y_i - a x_i - b) = 0 \iff$$

$$\sum_{i=1}^n y_i = a \sum_{i=1}^n x_i + n b \iff \bar{y} = a \bar{x} + b$$

Bigarren ekuaziotik,  $b = \bar{y} - a \bar{x}$  lortzen da eta lehen ekuazioan ordezkaturaz,

$$\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i^2 + (\bar{y} - a \bar{x}) \sum_{i=1}^n x_i = a \sum_{i=1}^n x_i^2 + \bar{y} \sum_{i=1}^n x_i - a \bar{x} \sum_{i=1}^n x_i$$

$$= a(\sum_{i=1}^n x_i^2 - n \bar{x}^2) + \bar{y} \sum_{i=1}^n x_i$$

Orduan,

$$a = \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{n s_x^2}$$

$$= \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{y} \bar{x}}{s_x^2} = \frac{s_{xy}}{s_x^2}$$

non,  $s_{xy} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{y} \bar{x}$  behaturiko balioen izeneko *kobariantza* den eta  $s_x^2$ ,  $x_i$  balioen bariantza.

Beraz,  $X$  gaineko,  $Y$  aldagaiaren erregresio zuzena, behaturiko puntuei hobereen doitzen zaiena, ondokoa izango da,

$$\tilde{Y} = a X + b = \frac{s_{xy}}{s_x^2} X + \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}$$

$$= \frac{s_{xy}}{s_x^2} (X - \bar{x}) + \bar{y}$$

edo gauza bera dena,

$$\tilde{Y} - \bar{y} = \frac{s_{xy}}{s_x^2}(X - \bar{x})$$

**Adibidea 6** Demagun baktería landaketa batean, denbora pasa ahala, bolumen unitateko ondoko baktería kopurua, ehunekotan neurturik, dauzkagula,

$X \equiv$ denbora ordutan	0	1	2	3	4	5
$Y \equiv$ baktería kopurua ehunekotan	12	19	23	34	56	62

Kalkula dezagun  $X$  eta  $Y$  aldagaien erregresio linealak,  $X$  gaineko,  $Y$ -ren erregresio zuzena eta  $Y$  gaineko,  $X$ -ena, eta estima dezagun baktería kopurua 6 ordu pasa ondoren eta pasatako denbora 10.000 baktería dagoenean.

Emandako datuekin, ondoko taula eraiki daiteke:

	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
	0	12	0	144	0
	1	19	1	361	19
	2	23	4	529	46
	3	34	9	1156	102
	4	56	16	3136	224
	5	62	25	3844	310
Guztira	15	206	55	9170	701

Beraz,

$$\bar{x} = \frac{5}{2}, \bar{y} = \frac{103}{3}, s_{xy} = \frac{\sum x_i y_i}{n} - \bar{y} \bar{x} = 31, s_x^2 = \frac{35}{12}, \text{ eta } s_y^2 = \frac{3146}{9}.$$

Orduan,  $X$  gaineko,  $Y$ -ren erregresio zuzena,

$$\tilde{Y} - \bar{y} = \frac{s_{xy}}{s_x^2}(X - \bar{x}) \iff \tilde{Y} - \frac{103}{3} = \frac{31}{\frac{35}{12}}(X - \frac{5}{2}) \iff$$

$$\tilde{Y} = \frac{372}{35}X + \frac{163}{21} \simeq 10,63X + 7,76$$

$$X = 6 \text{ denean, } \tilde{Y} = \frac{1073}{15} \simeq 71,53$$

Eta,  $Y$  gaineko,  $X$ -en erregresio zuzena,

$$\tilde{X} - \bar{x} = \frac{s_{xy}}{s_y^2}(Y - \bar{y}) \iff \tilde{X} - \frac{5}{2} = \frac{31}{\frac{3146}{9}}(Y - \frac{103}{3}) \iff$$

$$\tilde{X} = \frac{279}{3146}Y + \frac{857}{1573} \simeq 0,089Y - 0,54$$

$Y = 100$  denean,  $\tilde{X} = \frac{13093}{1573} \simeq 8,30$  (8 ordu eta ia 20 minutu)  
Horrez gain, Pearson-en koefizienteak,

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{31}{\sqrt{\frac{35}{12}} \sqrt{\frac{3146}{9}}} \simeq 0,9707$$

esaten digu,  $X$  eta  $Y$  aldagaien arteko korrelazioa oso altua dela.