Estatistika eta probabilitatea

HASTAPENAK

Itziar Irigoien
Jon Vadillo
Itsaso Rodríguez
Usue Mori
Borja Calvo



EHU

CIP. Unibertsitateko Biblioteka

Estatistika eta probabilitatea [Recurso electrónico]: hastapenak / Itziar Irigoien…[et al.]. – Datos. – [Leioa] : Servicio Editorial de la Universidad del País Vasco = Euskal Herriko Unibertsitateko Argitalpen Zerbitzua, [2026]. – 1 recurso en línea: PDF (216 p.). – (Unibertsitateko Eskuliburuak = Manuales Universitarios)
Modo de acceso: World Wide Web
ISBN: 978-84-1319-815-6

1. Estadística matemática. 2. Probabilidades. I. Irigoien Garbizu, Itziar, coaut.

(0.034)519.2

© Servicio Editorial de la Universidad del País Vasco
Euskal Herriko Unibertsitateko Argitalpen Zerbitzua

Argitalpen Zerbitzua • Servicio Editorial
argitaletxea@ehu.eus • 94 601 2227
Biblioteka eraikuntza, 1. solairua
Sarriena Auzoa, z/g. 48940 Leioa
www.ehu.eus/argitalpenak

ISBN: 978-84-1319-815-6

Gaien aurkibidea

Hitzaurrea **7**

1 Estatistika eta Probabilitaterako sarrera **9**
1.1 Sarrera laburra 9
1.2 Oinarrizko definizioak Estatistikan 12
1.3 Azterketa estatistikoen sailkapenak 13
1.4 Ezaugarria eta modalitateak. Aldagai estatistikoa 13

2 Estatistika deskribatzailea: maiztasunen banaketak eta adierazpen grafikoak **15**
2.1 Maiztasun-taulak 15
2.2 Adierazpen grafikoak 19

3 Estatistika deskribatzailea: joera zentral, sakabanapen eta posizio estatistikoak **33**
3.1 Joera zentraleko estatistikoak 34
3.2 Posizio estatistikoak 40
3.3 Sakabanapen estatistikoak 43
3.4 Batezbestekoaren eta desbideratze estandarraren erabilpen bateratua 49
3.5 Kutxa-diagrama 52

4 Estatistika deskribatzailea: bi dimentsiotako banaketak. Korrelazioa eta erregresioa **53**
4.1 Sarrera 53
4.2 Bi aldagai kuantitatiboren arteko erlazioa. Korrelazioa eta erregresio-zuzena 54

5 Probabilitate-kalkulua: oinarrizko probabilitate-teoria 71
5.1 Sarrera . . . 71
5.2 Gertaerak . . . 72
5.3 Probabilitate-funtzioa . . . 75
5.4 Nola finkatu probabilitateen balioak. Zorizko eredu klasikoa . . . 78
5.5 Probabilitate baldintzatua . . . 80
5.6 Bayes-en teorema . . . 83
5.7 Independentzia . . . 86

6 Probabilitate-kalkulua: zorizko aldagaiak eta probabilitate-banaketak 93
6.1 Zorizko aldagaiaren ideia orokorra . . . 93
6.2 Probabilitate-legea eta dentsitate-funtzioa . . . 95
6.3 Banaketa-funtzioa . . . 97
6.4 Itxaropena eta bariantza . . . 99
6.5 Hainbat banaketa ezagun . . . 103

7 Probabilitate-kalkulua: laginketa, simulazioa eta limitearen teoremak 119
7.1 Zenbaki handien legea . . . 119
7.2 Limitearen teorema zentrala . . . 120

8 Inferentzia estatistikoa: estimazioa 123
8.1 Sarrera . . . 123
8.2 Populazioa eta lagina . . . 125
8.3 Estimatzaileak eta beren propietateak . . . 125
8.4 Egoera arruntak . . . 134
8.5 Konfiantza-tarteak . . . 139
8.6 Bootstrap . . . 144

9 Inferentzia estatistikoa: hipotesi-probak 149
9.1 Sarrera . . . 149
9.2 Oinarrizko definizio eta kontzeptuak . . . 151
9.3 Hipotesi-proba egiteko urratsak . . . 152
9.4 p-balioen inguruan hainbat aipamen . . . 153
9.5 Lagin bakarrerako hipotesi-proba arruntenak . . . 154
9.6 Hipotesi-proben erabilerarako adibideak . . . 156
9.7 Bi laginetarako hipotesi-proba arruntenak . . . 160
9.8 Bi laginetarako hipotesi-proben erabilerarako adibideak . . . 162

10 Ariketak 165
1. eta 2. gaiak . . . 165
3. gaia . . . 170
4. gaia . . . 176
5. gaia . . . 178
6. eta 7. gaiak . . . 186

8. gaia 199
9. gaia 203

A Banaketa-taulak **209**
Banaketa binomiala 209
Poissonen banaketa 211
Banaketa normala 212
Studenten t-banaketa 213
Khi-karratua banaketa 214

Bibliografia **215**

Hitzaurrea

Euskal Herriko Unibertsitateko (EHU) Informatika Fakultatean *Informatika Ingeniaritzako* eta *Adimen Artifizialeko* graduko ikasketak eskaintzen dira. Bi gradu horietako ikasketa-planek *Estatistikako Metodoak Ingeniaritzan* irakasgaia jasotzen dute. Liburuxka hau egin dugunak aipatutako irakasgaiko irakasleak gara, eta ikasgelan jorratzen diren edukien osagarri izateko helburuarekin osatutako eskuliburua da hau; ikasleari kontzeptu teorikoen ulermena eta intuizioa hobetzeko lagungarri izateko asmoarekin egina. Kontzeptu horiek, egoki ikusi den kasuetan, R programazio-lengoaiako kode laburrarekin osatuta daude.

Gaur egun irakasle garenon eskutik dator eskuliburua, baina, bidea ibilian egiten delako, aurretik lankide izandako irakasleek ereindako aletik ere badu. Bereziki eskerrak eman nahi dizkiogu Yosu Yurramendi Mendizabali, urte askotako lankideari.

Liburuxka honetan, gaiaren sarrera orokorra (1. gaia) eta estatistika eta probabilitate-teoriari buruzko oinarrizko gaiak topatuko ditugu. Hiru atal nagusietan antolatu dira gai horiek:

1. atala. Estatistika deskribatzailea (2-4 gaiak)
2. atala. Probabilitate teoria (5-6 gaiak)
3. atala. Inferentzia estatistikoa (7-9 gaiak)

Halaber, liburuxkaren amaieran kontzeptuak lantzeko ariketa-zerrenda bat gehitu da, aipatutako ataletan eta dagozkien azpiataletan edo gaietan antolaturik.

1

Estatistika eta Probabilitaterako sarrera

1.1 Sarrera laburra

Bai *estatistika* eta bai *probabilitatea* oso hitz arruntak ditugu gure eguneroko jardunean, bereziki ziurgabetasuna eta datuen interpretazioa tartean daudenean. Izan ere, bizitzari eta munduari buruzko informazioa ematen diguten bitartekariek (prentsa, irrati, telebista, 'podcast'…) egunero azaltzen dizkigute estatistikak eta probabilitatean oinarritutako iragarpenak, informazio mota arrunt gisa. Horretaz gain, gure eguneroko elkarrizketa, iritzi eta gogoetetan ere sarritan tarteko ditugu datu estatistikoak eta probabilitatezko arrazoiketak. Kultura edo ezagutza orokorra halako informazioetan blai eginda dagoela ere esan daiteke. Gaur egungo gizartean estatistikak nonahi (eta nolanahi) erabiltzen baldin badira, beharrezkoa da pertsona orok gutxieneko ulermen-maila izatea, eta are beharrezkoagoa formakuntza teknikoko profesionalarentzat, teknikaria den heinean.

Ildo horretatik jarraituz, ikasketa teknikoetan graduatutako pertsonak, gainera, problematika jakin bati buruzko estatistikak antolatzen jakin beharko luke, gertakizunen probabilitateak balioztatzen, eta horien guztien atzean dagoen arrazoitzeko modua menperatzen.

Informazio estatistikoaren aurrean, askok honako bi jarraretako hauetako bat hartzen dute: bere mesfidantza sistematikoa ala zenbakien lotura miresgarriekiko fede hutsa. Halako bi muturretan besterik gabe ez erortzeko, zer hobeago informazio estatistikoa zertan datzan jakitea baino. Informazio estatistikoa informazio landua da, *estatistika* eta *probabilitatea* gaiek eskaintzen dituzten teoria eta tekniken bitartez landutakoa, hain zuzen.

Bi gaiak ezagutzea beharrezkoa egiten zaio ikasketa teknikoko graduatuari. Izan ere, har ditzagun informatika eta adimen artifizialarekin erlazionatuta dauden hainbat problema:

- Zerbitzarira iritsi den mezua SPAM den ala ez erabakitzea.
- Robotak «ikusi» duen irudian *larrialdi-irteera* ikurra badagoen identifikatzea.
- Zaintza-sistemak inguruan pertsona baten mugimenduaren segimendua egitea.
- Erabiltzailearentzat interesgarriak izan daitezkeen dokumentuak bilatzea Interneten.
- Erabiltzaileak emandako desbribapen bati erantzuten dion irudia sortzea.

Problema horiek ebazteko teknikak asko izan daitezke, eta arloko profesionalak teknika horiek ezagutu, erabili eta berrituko baditu, beharrezkoa zaio estatistikako eta probabilitateko oinarrizko kontzeptuak menperatzea. Izan ere, oinarrizko kontzeptu horiek ongi menperatzeak emango dio teknika berriak ulertzeko bidea. Horretan jardungo dugu ikasturte honetan. Problema horiek *adimen artifiziala*, *datu-analisia*, *ikasketa automatikoa* eta *datu-meatzaritza* arloetan txertaturik daude.

Horretaz gain, graduatua bere lanean diharduela, ziur aski beste arloko profesionalekin egingo du topo (sendagilea, ekonomilaria, soziologoa…), ezagutza-arlo gehienak jota baitaude teknologia informatikoarekin eta digitalarekin. Informazio estatistikoak problema askori aurre egiteko baliotsuak suertatu direnez gero, estatistika-metodoak garrantzi handikoak dira gaur egun ezagutza-arlo gehienetan. Maiz, ordea, beste arloetako espezialista horiek ez dute behar adina jakiten datu multzo handi bati etekin on bat ateratzeko edota informazio estatistikoa eskuratzeko garaian. Aplikazio estatistiko-informatiko bat nolanahi erabiltzea baino, aproposagoa izan daiteke estatistikaz eta probabilitatez jantzita dagoen profesional baten lankidetza gogoan hartzea. Beraz, estatistikaren eta probabilitatearen oinarrizko printzipioak ulertzea funtsezkoa da bai informatika-ingeniariarentzat eta bai adimen artifizialeko graduatuarentzat, beren jardun profesionalean ezinbesteko tresnak direlako.

Laburbilduz, estatistika eta probabilitatea ez dira gai bereziak, edonori dagozkionak baizik. Are gehiago, esperimentuan oinarritzen diren eginkizunetarako lagun paregabeak dira biak, eta, horrexegatik, ikasketa teknikoko graduatuak ezinbestez ezagutu behar lituzke estatistika-metodoak eta beren gutxieneko erabilpena.

Ikasketa teknikoko graduatuak zergatik menperatu behar dituen gutxieneko estatistika-metodoak arrazoitu dugu, eta orain ikus dezagun zein definizio formal topatzen ditugun *estatistika*rako eta *probabilitate*rako. Horrela, laburpen zehatz bat izateko *Elhuyar zientzia eta teknologiaren hiztegi entziklopediko*ra joko dugu. Bertan, honela dio:

- **Estatistika**

  Unitate multzo edo pertsona multzo bati dagozkion zenbakizko datuak biltzen, sailkatzen eta aztertzen dituen zientzia; delako zenbakizko emaitzak berak. Erabat ezagunak ez diren gertaera kolektiboen zenbakizko azterketa

egiteko analisi matematikoa (eragiketak, teknikak, eta metodoak) da estatistika.

(*) Ikusi, hurrengo estekan, entziklopedian bertan lotua duen artikulua: https://zthiztegia.elhuyar.eus/terminoa/eu/estatistika.

Estatistikaren barnean, askotan egiten da honako bereizketa hau:

- **Estatistika deskribatzailea**

  Populazio bati buruzko neurketak bildu, sinplifikatu nahiz laburtzen dituzten tekniken multzoa.

- **Estatistika inferentziala**

  Populazio baten ezaugarri kualitatibo zein kuantitatibo ezezagunak ulertzeko metodo matematikoak erabiltzea, haren laginetako datuak teknikoki maneiatuta, dela estimazio bidez, dela hipotesiak probatuz. Kasu partikularretatik abiatuta, maila orokorreko ondorioak ateratzea du helburua. Ondorio horietatik abiatuta, berriz, iragarpenak egin daitezke. Halaber, ondorioen adierazgarritasuna eta errore-gradua ebaluatu ohi dira probabilitate-kalkuluen bitartez; hartara, erabakiak ondorio horietan oinarrituta hartzen badira, oker egoteko probabilitatea ere neur daiteke.

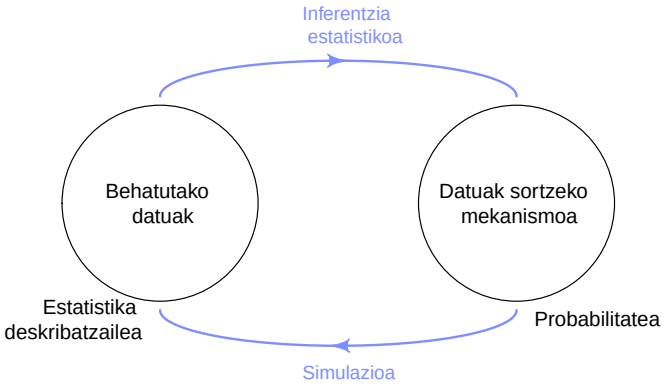
- **Probabilitatea**

  Estatistikan, gertaera jakin bat gauzatu dadin dagoen aukeraren neurria. Fenomeno edo saiakuntza batean gertaera jakin bat gauzatzen den kasuen (aldeko kasuen) eta beha daitezkeen gertaera edo kasu guztien arteko zatidura da. 1 baliotik (gertaera beti gauzatzen bada) 0 baliora (inoiz ere gauzatzen ez bada) bitartean alda daiteke.

  (*) Ikusi, hurrengo estekan, entziklopedian bertan lotua duen artikulua: https://zthiztegia.elhuyar.eus/terminoa/eu/probabilitate

Jatorriz bi gaiak desberdinak badira ere, zalantzarik gabe, bien arteko «loturak» hainbat arazori aurre egiteko tresna eskaintzen digu (Ikus 1.1. irudia). Batetik, estatistika atalean aipatutako «datuak» ditugu; bestetik, probabilitatea atalean, probabilitate-ereduak. Eredu horiek *datuak sortzeko mekanismo* eran ikus ditzakegu; izan ere, probabilitate-eredu bakoitzaren arabera, erraza da datu mota batzuk gertatzea eta, aldi berean zaila bestelakoak gertatzea. Horrela, datu batzuk probabilitate-ereduekin lotzeko aukera ezartzen da. Datuak ezagutuz, datu horiek orokortu ditzakeen probabilitate-eredua identifikatzeko bideari *inferentzia estatistikoa* esaten diogu. Alderantzizko bidean, probabilitate-eredu bat finkaturik datuak sortzeari, berriz, *simulazioa*.

Atal honetako gure helburua da oinarrizko *estatistika deskribatzailean* murgiltzea. Irakasgaian aurrerago murgilduko gara *probabilitatean*, eta irakasgaiaren azken zatian *estatistika inferentzialean* jardungo dugu.



1.1 irudia. Estatistika eta Probabilitatea kokatzeko diagrama.

1.2 Oinarrizko definizioak Estatistikan

Helburu jakin baterako informazioa duten datuak biltzea erabaki denean, komeni da honako kontzeptu hauek argi edukitzea:

- Azterketa estatistikoaren helburu den elementuen multzoari *populazioa* esaten zaio.
- Populazioko elementu bakoitzari *unitate estatistikoa* edo *alea* esaten zaio.
- Populazioaren azpimultzoari *lagina* esaten zaio. *Lagin-tamaina* laginaren elementu kopurua da, eta n letraren bidez adieraziko dugu. Populazio osoa aztertzea zaila edo ezinezkoa denean, lagina aukeratzen da.
- Populazioko elementuen gainean bildutako informazioari *ezaugarria* edo *aldagaia* esaten zaio. Ezaugarri bakoitzak era zehatzean definiturik egon behar du.

1.1 adibidea. Demagun Informatika Fakultateko ikasleen ohiturak zein diren jakin nahi duela Dekanotza taldeak. Bildu nahi duen informazioa jasotzeko, beheko galdetegia prestatu du.

1. *Zein da zure adina?*
2. *Zein da zure altuera (cm-tan)?*
3. *Ikasgelako lan-orduez gain, zenbat orduz lan egin beharko duzula astean uste duzu?*
4. *Ikasketekin batera, ordainduriko lanen bat egingo al duzu (B: bai, E: ez)?*
5. *Zure bizilekutik ikasgelara iristeko zenbat denbora behar duzu (minututan)?*
6. *Nola etortzen zara ikasgelara (O: oinez, B: bizikletaz, A: autobusez, X: bestela)?*
7. *Batez beste, egunean zenbat minutuz erabiltzen duzu Internet?*

Azterketa honen populazioa Informatika Fakultateko ikasleek osatzen dute. Galdetegia erantzungo duen ikasle taldeak osatuko du lagina. Horrela, ikasleek erantzungo duten galdera bakoitza ezaugarri bat dela esango dugu. Galderen erantzunetan izango dugu bildu nahi den informazioa.

1.3 Azterketa estatistikoen sailkapenak

Gauzatuko den azterketa estatistikoa hainbat eratara sailka daiteke. Batetik, *ale kopuruaren* arabera sailka daiteke, eta, bestetik, *ezaugarri kopuruaren* arabera.

1. **Ale kopuruaren araberako sailkapena**.

Azterketa estatistikoan populazioko elementu guztiak sartzen badira, *errolda* gauzatu dela esaten da. Aldiz, populazioko ale batzuk bakarrik aztertzen badira, azterketa estatistikoa *laginean* oinarritzen dela. Laginak baliozko informazioa bil dezan, funtsezkoa da lagina bera ongi aukeratzea. Horretaz arduratzen dira *laginketa* eta *esperimentuen diseinua* estatistikako arloak.

Azpimarratzekoa da azterketa estatistiko gehienak laginean oinarritutakoak direla, eta, gainera, egoera erreal askotan ez dagoela laginean oinarritutako azterketa burutzea beste aukerarik.

2. **Ezaugarri kopuruaren araberako sailkapena**.

Azterketa estatistikoan ezaugarri bakarra aztertzen bada, *aldagai bakarreko* azterketa esaten zaio. Ezaugarri bat baino gehiago batera aztertzen badira, *aldagai anitzeko* azterketa esaten zaio.

Oro har, azterketa interesgarrienak aldagai anitzekoak dira, baina, azterketa horiek egin ahal izateko, lehenengo ezaugarri bakar baten azterketa nola egin menperatu behar da. Aurtengo gure helburua izanfo da, lehenik eta behin, ezaugarri bakar baten azterketa nola egin, eta, ondoren, bi ezaugarriren azterketara hedatuko dugu ikasitakoa.

1.4 Ezaugarria eta modalitateak. Aldagai estatistikoa

Ezaugarriak bi multzo nagusitan banatzen dira:

1. **Kuantitatiboa**. Ezaugarriak aurkez ditzakeen egoerak zenbakiz neurgarriak direnean esaten zaio. *Aldagai estatistikoa* ere esaten zaio. Horien artean honako hauek bereizten dira:
   - *Diskretua:* aldagaiak hartu ditzakeen balioen multzoa finitua ala zenbagarria da.
   - *Jarraitua:* aldagaiak hartu ditzakeen balioen multzoak gutxienez (a, b) itxurako tarte erreala barneratzen du.

2. **Kualitatiboa**. Ezaugarriak aurkez ditzakeen egoerak ez-zenbakizkoak dira.

Ezaugarriak aurkez ditzakeen egoera ezberdinak *modalitateak* dira. Ezaugarriren baten modalitateak definitu behar direnean, modalitateak ongi definituta egoteko honako bi baldintza hauek bete behar dituzte:

1. **Zehaztasuna:** ezaugarriak aurkez ditzakeen egoera guztiak agertu behar dira.
2. **Bakartasuna:** ale bakoitzak modalitate bakarra aurkez dezake.

Bi baldintza horien esanahia erraz ulertuko dugu (1.1) adibideko 6. galderara itzuli eta aipatutako baldintzak betetzen ez diren moduari begiratzen badiogu. Adibidez, demagun galdera eta aurkezten diren modalitateak honako hauek direla:

6.a. *Nola etortzen zara ikasgelara (O: oinez, B: bizikletaz)?*

6.b. *Nola etortzen zara ikasgelara (O: oinez, B: bizikletaz edo oinez, A: autobusez, X: bestela)?*

Lehen kasuan (6.a), zehaztasuna falta dugu: autobusez datorrenarentzat ez dago modalitaterik definituta, ezingo luke dagokion erantzun zuzena eman. Bigarren kasuan (6.b), bakartasuna: oinez datorrenarentzat modalitate bat baino gehiago daude, ezingo luke erantzun bakarra eman.



Sakontzeko

1. Pentsa ezazu egoera bat, non derrigorrez laginaren gainean egin behar baita azterketa estatistikoa. Arrazoitu ezazu zergatik.
2. Esan ezazu (1.1) adibideko galdera bakoitza zein motakoa den.
3. Demagun honako galdera hau erantzun behar duzula:
   PC ordenagailuen errendimendua zeren araberakoa da?
   Zer egin dezakezu galdera hori erantzuteko? Zein ordenagailu motarentzat erantzun nahi duzu galdera hori? Jar ezazu egoera horretan zure burua, eta erabaki ezazu zein izango diren:
   - Populazioa
   - Ale/unitate estatistikoa
   - Lagina
   - Aldagai estatistikoak

2

Estatistika deskribatzailea: maiztasunen banaketak eta adierazpen grafikoak

Gogora dezagun gure helburu nagusia datuak laburbiltzea dela. Hau da, datuetan dagoen informazioa modu erraz eta ulergarrian adieraztea. Lehenengo, tauletan laburbilduko ditugu datuak, eta, ondoren, grafikoetan erakutsiko. Orrialde hauetan, ezaugarri bakar bat dugun kasura mugatuko gara.

2.1 Maiztasun-taulak

Demagun ezaugarri batentzat n tamainako laginean behatutako balioak ditugula. Datu multzo hori ordenatzea izango da lehenengo lana. Horretarako, maiztasun-taula edo taula estatistikoa eraikiko dugu.

Maiztasun-taula ondoren aipatzen diren zutabeekin eraikitzen da:

1. **Ezaugarriaren behatutako balioak:** $x_1, \dots, x_k$. Ezaugarria kuantitatiboa bada, balio horiek txikienetik handienera ordenatuta.
2. **Maiztasun absolutuak:** $n_1, \dots, n_k$, non n_i zenbakiak x_i balioari dagokion lagineko ale kopurua adierazten baitu ($i = 1, \dots, k$).
3. **Maiztasun absolutu metatuak:** $N_1, \dots, N_k$, non:

$$N_i = n_1 + \dots + n_i = \sum_{l=1}^{i} n_l\,, \quad i = 1, \dots, k.$$

4. **Maiztasun erlatiboak:** $f_1, \dots, f_k$, non:

$$f_i = n_i/n\,, \quad i = 1, \dots, k.$$

5. **Maiztasun erlatibo metatuak:** $F_1, \ldots, F_k$, non:

$$F_i = N_i/n\,, \quad i = 1, \ldots, k.$$

Definizio horietatik, honako propietate hauek ondoriozta daitezke:

2.1 propietatea.

$0 \leq n_i \leq n\,; \quad 0 \leq f_i \leq 1; \quad i = 1, \ldots, k.$

2.2 propietatea.

$\sum_{i=1}^{k} n_i = n; \quad N_k = n\,; \quad \sum_{i=1}^{k} f_i = 1\,; \quad F_k = 1\,.$

2.1 adibidea. Azken astean zehar, R programaren instalaziorako deskargak zein plataformatarako izan diren aztertu da. Hauek dira lorturiko erantzunak (jasotako datuen zati bat bakarrik ikus dezakezu, bestela oso luzea litzatekeelako):

Linux	Linux	Windows	Windows	Windows
Linux	Windows	MacOS X	Windows	Linux
...	...	...	...	...

Adibide honetan argi ikusten da datuak laburbiltzeko beharra. Datu gordinak banan-banan jartzeko bi orrialde beharko genituzke, eta, gainera, nekez geureganatuko genuke bertan dagoen informazioa. Aldiz, maiztasun-taulak berehala erakusten digu datuetan dagoen informazioa.

Maiztasun-taula

Plataforma (x_i)	n_i	f_i
Linux	309	0.31[a]
MacOS X	84	0.08
Windows	607	0.61
Guztira	1000	1

[a] Euskaraz, digitu hamartarren marka ","-rekin adierazi ohi da, Euskaltzaindiaren arauari jarraituz. Kasu honetan, 0,31 litzateke. Eskuliburu honetan ordea R-ko agindu eta emaitzak ditugunez, notazio bateratua erabili dugu, hamartarren marka "."-rekin adieraziz. Era berean, milakoen marka (".") ez dugu erabili.

2.2 adibidea. Inkesta batean seme-alaba kopurua zein zen galdetu da. Hauek dira lorturiko erantzunak:

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

Goiko aldeko zenbaki sorta zuzenean ikusita, nekez jakingo dugu nola banatzen den sema-alaba kopurua. Seme-alaba kopurua nola banatzen den errazago ikusteko, informazio hori maiztasun-taula batean laburbiltuko dugu.

Maiztasun-taula

x_i	n_i	N_i	f_i	F_i
0	4	4	0.100	0.100
1	17	21	0.425	0.525
2	10	31	0.250	0.775
3	6	37	0.150	0.925
4	2	39	0.050	0.975
5	0	39	0.000	0.975
6	1	40	0.025	1
Guztira	40		1	



Pentsa ezazu...

- Zer adierazten du maiztasun-taulako zutabe bakoitzak? Zergatik uste duzu ez direla erakutsi maiztasun metatuen zutabeak lehen adibidean?
- Begira iezaiozu momentu batez jarraian duzun 2.3 adibideari. Uste duzu aurreko adibidean bezalako maiztasun-taula eraikitzea egokia dela gure helbururako? Zergatik? Nolako aldagaiak dira adibide bakoitzekoak?

2.3 adibidea. eBay webgunean azken hilabetean *Smartphone* produktuak enkantean izan dituen salmenta-prezioak (eurotan).

250	172	180	419.5	148	152	264	177.5
137.5	175	107	399.5	220	134.9	142	134
224.5	142	132	150	212.5	142	194	132.5
61	125	145.5					

Egoera honetan, egokia da maiztasun-taula eraiki aurretik behatutako balioak tartekatzea. Horretarako, lehenengo hartu beharreko erabakia izango da *zenbat tartetan* multzokatuko ditugun balioak. Baldin eta tarte askotan egitea erabakitzen badugu, ez dugu gure hasierako arazoa ebatziko; baina baldin eta oso tarte gutxitan multzokatzea erabakitzen badugu, bidean informazio asko galduko dugu. Erabaki hori hartzeko ez dago era zuzen bakarra, baina bi bideen arteko oreka eskaintzen du Sturges-en formulak. Formula horren arabera, k tartetan multzokatzeko honela aukeratuko dugu:

$$k \approx 1 + 3.322 log_{10} n,$$

n lagin-tamaina delarik.

2.3. Adibidearen ebazpena

Datuak multzokatuko ditugu Sturgesen formula erabiliz.

- $n = 27$
- Tarte kopurua Sturgesen formularen bidez:

$$k \approx 1 + 3.322 log_{10} n = 5.75 \quad \Rightarrow \quad k = 6$$

- Aldagaiaren heina:

$$\begin{aligned} R &= 419.5 - 61 \\ R &= 358.5 \end{aligned}$$

- Tarte bakoitzaren zabalera:

$$\begin{aligned} a_i &\approx 358.5/6 \\ a_i &= 60 \end{aligned}$$

- Tarteak osatu. Norberak erabakiko du zein izango diren tarteen mugak eta, noski, aukera asko dago horretarako. Komeni da zentzuzkoak izatea aukeratutako tarteen mugak aztertzen ari garen testuinguruan eta, batez ere, behatutako balio guztiak tarteren batean eta bakar batean barneratuak egotea.

  $l_1 = 60$ finkatzen badugu:

$$l_2 = 120 \quad l_3 = 180 \quad l_4 = 240 \quad l_5 = 300 \quad l_6 = 360 \quad l_7 = 420$$

Maiztasun-taula, datuak elkartuta

$[l_i, l_{i+1})$	n_i	N_i	f_i	F_i	x_i	a_i
$[60, 120)$	2	2	0.074	0.074	90	60
$[120, 180)$	16	18	0.593	0.667	150	60
$[180, 240)$	5	23	0.185	0.852	210	60
$[240, 300)$	2	25	0.074	0.926	270	60
$[300, 360)$	0	25	0	0.926	330	60
$[360, 420)$	2	27	0.074	1	390	60
Guztira	27		1			

Taula horretan jaso diren zutabe berriak hauek izan dira:

1. Eraiki diren tarteen mugak: $[l_i, l_{i+1}), \quad i = 1, \dots, k$. Aukerazkoa bada ere, irakasgaian zehar oro har $[\cdot\ , \cdot)$ itxurako tarteak eraikiko ditugu.

2. Tartearen *klase-marka*: $x_i, i = 1, \dots, k$.

$$x_i = \frac{l_i + l_{i+1}}{2},$$

   tartearen erdiko puntua tarte horretan biltzen diren balioen ordezkaria da.

3. Tartearen zabalera: $a_i, i = 1, \dots, k$.

$$a_i = l_{i+1} - l_i.$$

Elementu horiek zenbait grafiko eraikitzeko eta datuak laburbiltzeko erabiliko ditugu aurrerago.

2.2 Adierazpen grafikoak

Datuetan dagoen informazioa modu azkarrean erakusteko beste tresna oso egokiak izaten dira grafikoak. Grafikoak, ordea, ezin du edozein izan: egoera bakoitzean, ezaugarriaren izaeraren arabera grafiko mota batzuk ala beste batzuk izango dira

egokiak. Jarraian, 3 adibide topatuko dituzu (2.4, 2.5 eta 2.6 adibideak). Lehenengo, saiatu zaitez adibide bakoitzaren gainean hausnartzen; ondoren, ikus itzazu proposatzen diren grafikoak, eta, azkenik, azter ezazu grafiko bakoitza eta ikas ezazu nola eraikitzen den. Azken horretarako, hiru alderdi ikusiko dituzu. Bata, grafikoak zer daturen ezaugarri erakutsi nahi duen eta horren baitan nola eraikitzen den. Bigarrena, Rrekin nola eraiki daitekeen. Horretarako, erakusten den grafiko bakoitzaren azpian grafikoa erakitzeko Rko kodea duzu. Kode horrek erakusten den grafikoaren bertsio sinple bat berregingo du, ez erakusten den grafikoa xehetasun berdinekin. Azkenik, grafiko mota horren inguruko azalpenak ere jaso dira.

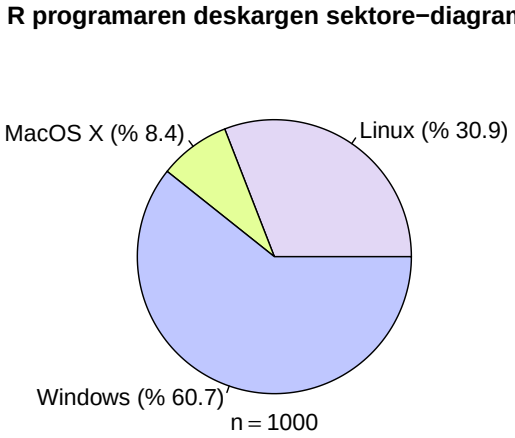
2.4 adibidea. Azken astean, R programaren instalaziorako deskargak zein plataformatarako izan diren aztertu da.

Plataforma	n_i
Linux	309
MacOS X	84
Windows	607
	$n = 1000$

Ezertan hasi aurretik, azter dezagun: Nolako ezaugarria da? Kasu honetan, aldagai kualitatiboa da, eta modalitate gutxi ditu. Hori horrela izanik, bi grafiko mota proposatzen dira:

1. Sektore-diagrama. Sektore-diagrama batean zirkuluak osotasuna adierazten du, eta modalitate bakoitzari sektore bat dagokio, **sektore bakoitzaren azalera dagokion modalitatearen maiztasunarekiko proportzionala** izanik. Diagrama mota hau erabili ohi da proportzioak edo datuen banaketa orokorra erakusteko, batez ere modalitate gutxi daudenean. Hala ere, antzeko azalera duten sektoreen arteko desberdintasunak konparatzea edo balio zehatzak antzematea zaila izan daiteke sektore-diagrametan. Horrelako kasuetan, egokiagoak izaten dira jarraian ikasiko ditugun barra-diagramak.

Aurreko 2.4 adibidearekin jarraituz, honako itxura hau izango du dagokion sektore-diagramak:



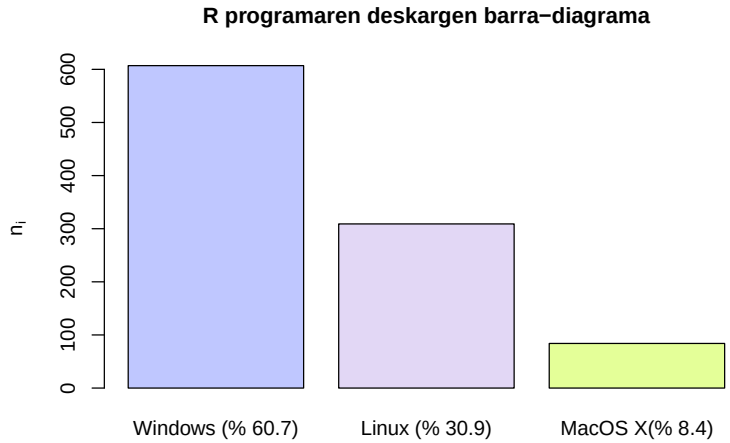
Azpiko kodearekin sektore-diagrama egin dezakezu. Erreparatu kode honekin ez dituzula koloreak aukeratu.

```
a <- c(309, 84, 607) # Maiztasunak jaso bektore batean
etiketak <- c("Linux", "MacOS X", "Windows") # Sektoreentzako
                                             # etiketak

pie(a, labels= etiketak,
    main="R programaren deskargen sektore-diagrama")
```

2. Barra-diagrama. Barra-diagrama batean barra bakoitzak modalitate bat adierazten du, eta **zutabearen altuerak dagokion modalitatearen maiztasuna (absolutua edo erlatiboa)**. Diagrama mota hau oso erabilia da datu kualitatiboak zein datu kuantitatibo diskretuak irudikatzeko, balio desberdinen arteko konparazio zuzen eta bisual bat ahalbidetzen duelako.

Berriro hartuz 2.4 adibidea, azpian dugu dagokion barra-diagrama:




```r
a <- c(309, 84, 607) # Maiztasunak jaso bektore batean
etiketak <- c("Linux", "MacOS X", "Windows") # Barrentzako
                                             # etiketak

barplot(a, names.arg=etiketak,
        main="R programaren deskargen barra-diagrama")
```

2.5 adibidea. Plataforma baten erabiltzaileen artean, azken astean zehar erabiltzaile bakoitza plataformara zenbat aldiz sartu den aztertu da. Hauek dira behaturiko balioak.

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

Berriro ere, grafikoak egiten hasi aurretik, datuen inguruan hausnartzea komeni zaigu: Plataformako zenbat erabiltzaile dago laginean? Asko badira, komeniko zaigu maiztasun-taula batean laburbiltzea. Eta, nolako ezaugarria da? Ezaugarri kuantitatibo diskretua izanik, barra-diagrama egokia izan daiteke datu horiek erakusteko. Ez da, ordea, aukera bakarra; jarraian dituzu proposatutakoak:

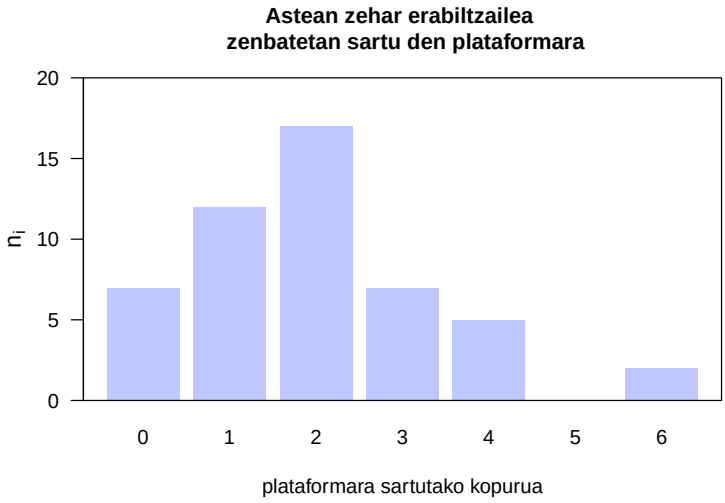
Aipatu bezala, 2.5 adibideko datuen grafikoak eraikitzeko, komeni da aurretik datu horiek maiztasun-taula batean antolatzea.

Maiztasun-taula

x_i	n_i	f_i	N_i	F_i
0	7	0.14	7	0.14
1	12	0.24	19	0.38
2	17	0.34	36	0.72
3	7	0.14	43	0.86
4	5	0.10	48	0.96
5	0	0.00	48	0.96
6	2	0.04	50	1

1. Barra-diagrama. Lehen ikusi dugu barra-diagramak nola adierazten diren. Aldagaia kuantitatibo diskretua denean ere egokia izaten da.

Eraiki berri dugun maiztasun-taulan oinarrituta, 2.5 adibideko datuak barra-diagraman adierazita.



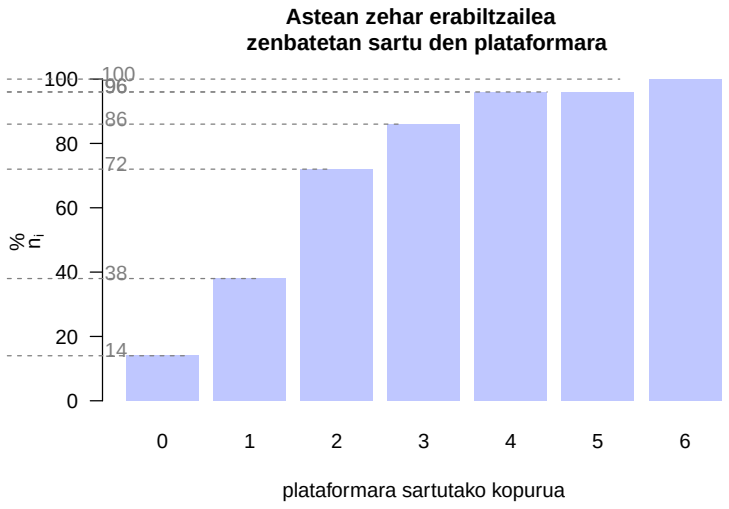
```
# Sartu zure datuak x objektuan
x <- c(2, 1, 2, 0, ...) # OSATU

#Kontatu zenbat aldiz agertzen den balio bakoitza
ni <- tabulate(x+1) # KONTUZ 0 balioarekin,
                    # ez du kontutan hartzen!!!

barplot(ni, names.arg=0:6,
        xlab="plataformara sartutako kopurua",
        ylab="Maiztasunak",
        main="Astean zehar erabiltzailea \n
              zenbatetan sartu den plataformara")
```

2. Barra-diagrama metatua. Barra-diagrama metatuan, barra bakoitzak modalitate bat adierazten du, eta zutabearen altuerak dagokion modalitatearen **maiztasun metatua** (absolutua edo erlatiboa). Diagrama mota hau soilik datu kuantitatiboak adierazteko da egokia.

2.5 adibideko datuak kuantitatibo diskretuak direnez, barra-diagrama metatuarekin erakuts ditzakegu.



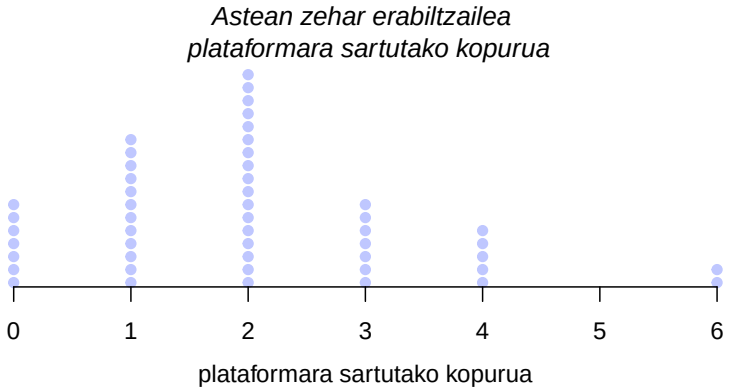
```r
# Sartu zure datuak x objektuan
x <- c(2, 1, 2, 0, ...) # OSATU
#Kontatu zenbat aldiz agertzen den balio bakoitza
ni <- tabulate(x+1) # KONTUZ 0 balioarekin,
                    # ez du kontutan hartzen!

Fi <- cumsum(ni)/sum(ni) # Maiztasun erlatibo metatuak

barplot(Fi, names.arg=0:6,
        xlab="plataformara sartutako kopurua",
        ylab="Proportzioak",
        main="Astean zehar erabiltzailea \n
              zenbatetan sartu den plataformara")
```

3. Puntu-diagrama. Puntu-diagrama ezaugarri kuantitatibotarako erabiltzen da eta **ale bakoitza irudikatzen da grafikoan**, datuak maiztasun-taula batean antolatu gabe. Ardatz horizontal batekin eskala ezartzen da, eta ale bakoitzaren balioa eskala horretan kokatu.

2.5 adibideko datuentzako puntu-diagrama.




```r
# Sartu zure datuak x objektuan
x <- c(2, 1, 2, 0, ...) # OSATU

stripchart(x, method="stack",
           xlab="plataformara sartutako kopurua",
           main="Astean zehar erabiltzailea \n
                 zenbatetan sartu den plataformara")
```

2.6 adibidea. e-salmentarekin zerikusia duten 100 enpresek azken urtean izandako irabaziak (milioi eurotan) hauek izan dira:

16.3	24.5	31.7	42.6	48.8	51.2	54.9	55.3	55.3	61.7
62.7	64.4	66.2	66.5	67.7	69.9	70.2	72.6	72.8	73.3
73.7	74.8	75.4	75.8	77.1	78.9	79.0	79.4	80.4	80.6
81.2	82.5	83.1	83.2	83.5	85.1	87.0	87.4	88.1	88.1
88.9	89.3	90.0	90.3	90.7	91.2	91.5	91.5	91.6	91.6
92.0	95.2	97.5	97.7	98.3	98.9	99.1	99.6	101.6	104.7
104.9	105.4	105.5	106.2	106.8	108.9	110.0	110.5	111.0	113.9
114.3	115.7	116.1	116.3	117.7	117.9	120.1	120.3	121.2	122.2
124.2	124.9	125.1	126.3	126.9	128.8	130.0	131.2	133.9	136.8
136.9	140.2	140.4	140.4	140.6	141.6	147.8	149.0	159.9	165.8

Berriro ere, ezertan hasi aurretik, egoera zein den pentsatu behar dugu: Zein da aztertu beharreko ezaugarria? Nolako ezaugarria da? *Irabaziak* badira, ezaugarri kuantitatibo eta jarraitua dugu. Datu horiek irudikatzeko, lehenengo maiztasun-taularen batean ordenatzea egokia da, baina ez da guztiz derrigorrezkoa. Ikus ditzagun datu horiek irudikatzeko hainbat aukera.

1. Zurtoin eta hosto grafikoa (stem-and-leaf). Zurtoin eta hosto grafikoan **ale bakoitza irudikatzen da**, datuak maiztasun-taula batean jaso gabe. Ale bakoitzari dagokion balioa (x_i) *zurtoin* eta *hosto* elementutan banatzen da. *Hostoa* x_i balioaren azken digitoa da. *Zurtoina*, gainontzeko digituekin osatutako zenbakia. Adibidez, $x_i = 16$ bada, *hostoa*(x_i)= 6, eta, ondorioz, *zurtoina*(x_i)=1. Zurtoin denak betikalki ordenatu eta aldamenean dagozkien hostoak jarriz osatzen da grafikoa. Azkenik, digitu hamartarren marka zein posiziotan kokatzen den adierazten da. Eskuz, grafikoa azkar egiteko modua eskaintzen du, baina ez da dotoreena.

Jarraian, 2.6 adibideko datuak zurtoin eta hosto grafikoetan irudikatuta daude.

```
The decimal point is 1 digit(s) to the right of the |

  1 | 6
  2 | 5
  3 | 2
  4 | 39
  5 | 1555
  6 | 234678
  7 | 0033345567999
  8 | 01133345778899
  9 | 001122222588899
 10 | 025556679
 11 | 0114466688
 12 | 0012455679
 13 | 01477
 14 | 0001289
 15 |
 16 | 06
```

```r
# Sartu zure datuak x objetuan
x <- c(16.3, 24.5, ...) # OSATU

# Zurtoin eta hosto grafikoa eraiki
stem(x, scale=2)
```

Oharra. Adibide honetan, grafikoa eraiki aurretik datuak batekoetara biribildu dira. Horrela, $x_1 = 16.3$ behaketa 16.0 baliora biribilduta adierazi da goiko zurtoin eta hosto grafikoan.

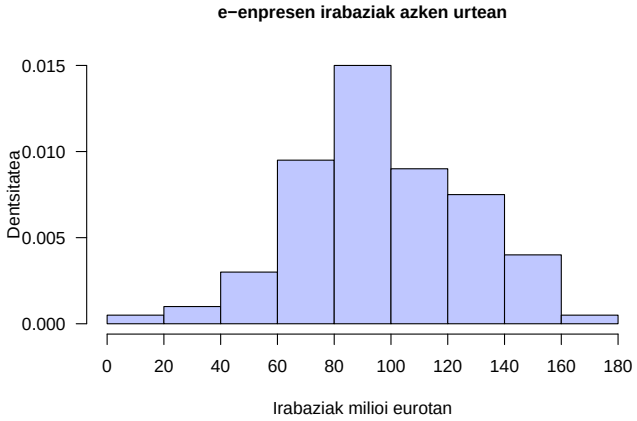
2. Histograma. Histograma batek badu barra-diagrama baten antza, baina baita desberdintasunak ere. Histogramak aldagai jarraituak adierazteko erabiliko ditugula, hori da desberdintasun nabarmenena. Horrela, ardatz horizontalean **aldagaia tarteetan banatuta** adieraziko da. Tarte bakoitzaren gainean eraikiko den **laukizuzenaren azalerak tarte horretan aldagai estatistikoak erakutsitako maiztasunekiko proportzionala** izan behar du.

Jarraian, 2.6 adibideko datuak bi maiztasun-taula desberdinen arabera laburbildu dira, taula bakoitzetik eraikitako histogramarekin batera.

Histograma (A)

Histograma maiztasun-taula honetan oinarrituta eraikita dago.

Irabaziak	n_i	n_i/a_i	Dentsitatea: $f_i/a_i = \frac{n_i}{a_i} \cdot \frac{1}{n}$
$[0, 20)$	1	0.05	0.0005
$[20, 40)$	2	0.10	0.0010
$[40, 60)$	6	0.30	0.0030
$[60, 80)$	19	0.95	0.0095
$[80, 100)$	30	1.50	0.0150
$[100, 120)$	18	0.90	0.0090
$[120, 140)$	15	0.75	0.0075
$[140, 160)$	8	0.40	0.0040
$[160, 180)$	1	0.05	0.0005
	100		



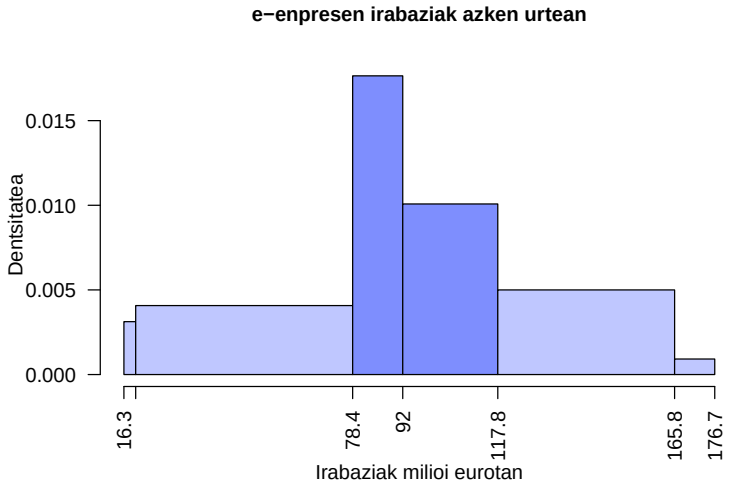

```r
# Sartu zure datuak x objektuan
x <- c(16.3, 24.5, ...) # OSATU

# Histograma eraiki
hist(x, freq=FALSE,
     xlab="Irabaziak milioi eurotan", ylab="Dentsitatea",
     main="e-enpresen irabaziak azken urtean")
```

Histograma (B)

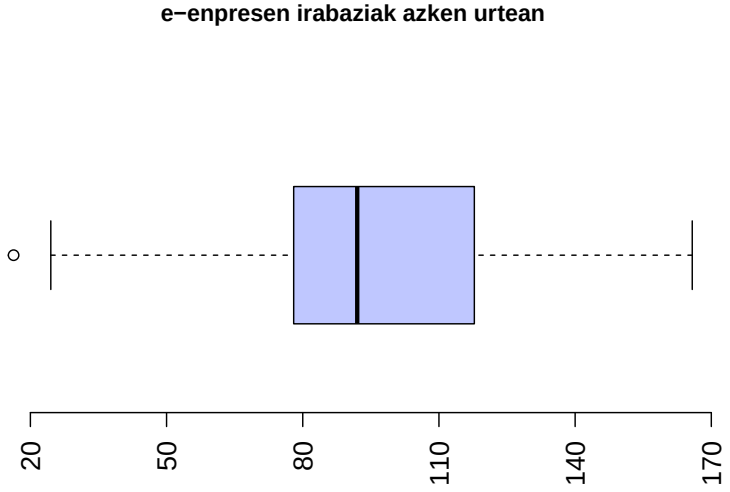
Histograma maiztasun-taula honetan oinarrituta eraikita dago.

Irabaziak	n_i	n_i/a_i	Dentsitatea: $f_i/a_i = n_i/a_i \times 1/n$
$[16.3, 19.5)$	1	0.31	0.0031
$[19.5, 78.4)$	24	0.41	0.0041
$[78.4, 91.8)$	25	1.87	0.0187
$[91.8, 117.8)$	25	0.96	0.0096
$[117.8, 165.8)$	24	0.50	0.0050
$[165.8, 176.7)$	1	0.09	0.0009
	$n = 100$		



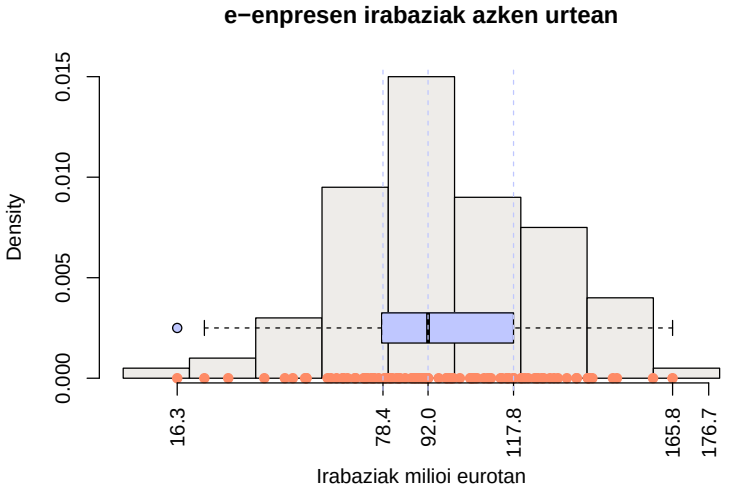
3. Kutxa-diagrama. Kutxa-diagrama aldagai kuantitatiboak, diskretuak zein jarraituak, adierazteko oso grafiko aproposa da. Hala ere, *kutxa-diagrama* eraikitzeko zenbait kontzeptu gehiago ikasi behar dituzu. Horregatik, irakasgaian aurrera egingo dugu, eta hurrengo gaiaren amaieran ikasiko dugu nola eraiki eta nola interpretatu grafiko mota hau (Ikus kutxa-diagrama eraikitzeko jarraibideak, 3.5 atala).

2.6 adibideko datuak kutxa-diagrama batean dituzu azpian. Baita kutxa-diagrama puntu-diagramarekin eta histogramarekin batera ere bigarren grafiko batean, interpretazioan laguntzeko.




```r
# Sartu zure datuak x objektuan
x <- c(16.3, 24.5, ...) # OSATU

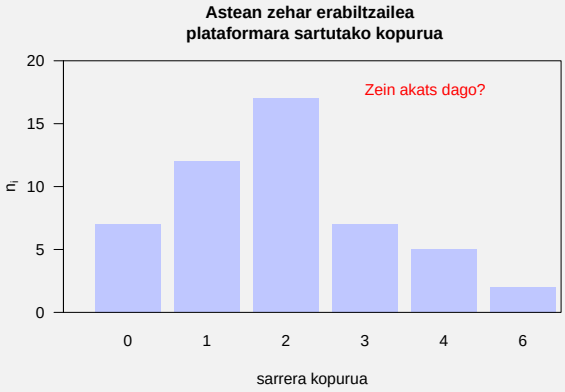
# Kutxa-diagrama eraiki
boxplot(x, horizontal=TRUE,
        xlab="Irabaziak milioi eurotan",
        main="e-enpresen irabaziak azken urtean")
```





Sakontzeko

- Atal honetako helburu nagusiak bi dira: bata, behatutako datuak taula egoki batean laburbiltzeko gai izatea, eta, bestea, datuak aurkezteko grafiko egokia zein den aukeratu eta eraikitzea. Eskuliburuaren amaieran jasotako 2. gaiko ariketa-zerrendako ariketak egin itzazu helburuak lortzeko.
- Gure plataformako erabiltzaileak astean zehar zenbat aldiz sartzen diren plataformara jaso da azpiko grafikoan. Zer akats ikusten duzu grafikoan?



- Demagun $n = 50$ tamainako laginean pertsonen altuera neurtu dela eta honako datu hauek lortu ditugula.

Altuera	n_i
$[1.40, 1.60)$	6
$[1.60, 1.70)$	22
$[1.70, 1.75)$	12
$[1.75, 1.80)$	7
$[1.80, 1.90)$	3
	$n = 50$

  Eraiki ezazu histograma, datu horiek laburbiltzeko taulak erakusten dituen tarteetan.
- Ikusitako grafiko mota bakoitza eraikitzen jakin behar duzu. Gainera, lagungarri izango zaizu ikasitakoa itxura honetako taula batean laburbiltzea:

Ezaugarri mota	Grafiko egokia	Oharrak
Kualitatiboa	⋮	⋮
Kuantitatiboa	⋮	⋮

3

Estatistika deskribatzailea: joera zentral, sakabanapen eta posizio estatistikoak

Populazioko lagin batetik lortutako datu multzoa dugunean, datu horiek duten informazioa laburbiltzea dugu helburu. Horretarako dira ikusi berri ditugun taula estatistikoa eta adierazpen grafikoak. Hala ere, gehienetan hori ez da nahikoa izaten, zenbakizko laburpena ere behar izaten dugu. Zenbakizko laburpen horiek *estatistikoen* bidez egingo ditugu. Beste era batera esanda, datuen ezaugarriren bat zenbaki batekin laburbiltzen duen horri *estatistiko* esaten zaio.

Alderdi edo ikuspegi desberdinetatik laburbildu behar ditugu datuak lortuko dugun deskribapena osatua izango bada, eta, oro har, estatistiko bakoitzak datuen alderdi bat laburbiltzen du. Adibidez, askotan jakin nahiko dugu zein «erdiko» balioren inguruan kokatzen den ezaugarria (*Joera zentrala*), eta interesgarria izango zaigu jakitea «erdiko» balio horren inguruan oso konzentraturik ala sabakanaturik azaltzen diren ezaugarriaren balioak (*Sakabanapena*). Halaber, ezaugarriari buruzko informazio erlatiboa eskainiko digute posizio jakinetan kokatzen diren datuek (*Posizio estatistikoak*). Badaude beste ikuspegi batzuk ere, baina guk ez ditugu hemen landuko. Azkenik, aipa dezagun ikuspegi batetik laburbiltzeko ditugun estatistikoak bat baino gehiago direla. Jarraian, ikuspegi bakoitza aztertzeko zein estatistiko erabiliko ditugun azalduko dugu.

Hemendik aurrera, aztergai dugun ezaugarria kuantitatiboa dela suposatuko dugu; izan ere, ikusiko ditugun estatistikoek ez dute zentzurik izango ezaugarri kualitatiboentzat.

3.1 Joera zentraleko estatistikoak

Joera zentraleko estatistikoek adierazten dute aldagaiak zein balioren inguruan hartzen dituen balioak. Bi aztertuko ditugu hemen:

- Batezbesteko aritmetikoa.
- Mediana.

3.1 definizioa. Laginean neurturiko X aldagai estatistikoaren *batezbesteko aritmetikoa* ($\bar{x}$) honela definitzen da:

- Datuak elkartuta badaude, hau da, $x_1, \dots, x_k$ behatutako balioen maiztasunak $n_1, \dots, n_k$ badira, hurrenez hurren:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i.$$

- Elkartu gabeko datuak badira, $x_1, \dots, x_n$:

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

3.1 adibidea. Inkesta batean seme-alaba kopurua zein zen galdetu da. Hauek dira lorturiko erantzunak:

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

Lagin honetan, zein da batez besteko seme-alaba kopurua?

Ebazpena

Aurretik eraikita zenuen maiztasun-taulan zutabe berri bat gehituko duzu kalkuluak errazteko:

x_i	n_i	N_i	f_i	F_i	$n_i \times x_i$
0	4	4	0.1	0.1	0
1	17	21	0.425	0.525	17
2	10	31	0.250	0.775	20
3	6	37	0.150	0.925	18
4	2	39	0.050	0.975	8
5	0	39	0	0.975	0
6	1	40	0.025	1	6
Guztira	40		1		69

Orduan,

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i \\ &= \frac{69}{40} \\ &= 1.725. \end{aligned}$$

Beraz, lagin honetan batezbeste 1.7 seme-alaba du bakoitzak.

Horixe bera egin dezakezu R erabiliz. Erreparatu kasu honetan, datuak elkartu gabe jaso direla.

```
# Behatutako balioak jaso objektu batean
x <- c(2, 2, 1, ...) # Osatu

# Lagin-tamaina
n <- length(x)
# Emaitza: 40

# Balio guztien batura
batura <- sum(x)
# Emaitza: 69

# Batezbestekoa:
batura/n
# Emaitza: 1.725
```



Sakontzeko

Itzul gaitezen dagoeneko ikusia dugun datu multzora:
eBay webgunean azken hilabetean *Smartphone* produktuak enkantean izan dituen salmenta-prezioak (eurotan).

250	172	180	419.5	148	152	264	177.5
137.5	175	107	399.5	220	134.9	142	134
224.5	142	132	150	212.5	142	194	132.5
61	125	145.5					

1. Kalkula ezazu zein izan den smartphoneak izan duen batezbesteko salmenta-prezioa.
2. Demagun, datu gordinak izan beharrean, honako taula honetan ordenaturik eman dizkigutela.

$[l_i, l_{i+1})$	n_i	N_i	f_i	F_i	x_i	a_i
$[60, 120)$	2	2	0.074	0.074	90	60
$[120, 180)$	16	18	0.593	0.667	150	60
$[180, 240)$	5	23	0.185	0.852	210	60
$[240, 300)$	2	25	0.074	0.926	270	60
$[300, 360)$	0	25	0	0.926	330	60
$[360, 420)$	2	27	0.074	1	390	60
Guztira	27		1			

   Zein batezbesteko balio lortuko duzu taulan oinarrituz?
3. Lortu dituzun bi balioak desberdinak izan dira. Zer iritzi duzu horri buruz?

Batezbestekoaren zenbait propietate estatistiko:

- Bai kalkulatzeko eta baita matematikoki maneiatzeko ere oso erraza da.
- Oztoporik handiena du lagin-fluktuazioekiko sentikorra izatea. Hau da, lagineko muturreko balioek eragina dute beregan.

3.2 adibidea. Izan bedi lagin hau: $x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 5, x_5 = 5, x_6 = 6, x_7 = 300$.

$$\begin{aligned} \bar{x} &= \frac{1}{7}(2+3+4+5+5+6+300) \\ &= 46.43 \end{aligned}$$

46.43 zenbakiak laginaren erdialdea ongi adierazten duela uste al duzu?

3.1 propietatea.

Izan bitez $x_1, \dots, x_k$ lagineko balioak eta $n_1, \dots, n_k$ dagozkion maiztasun absolutuak, hurrenez hurren. Orduan,

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

Frogapena.

$$\begin{aligned} \sum_{i=1}^{k} n_i(x_i - \bar{x}) &= \sum_{i=1}^{k} n_i x_i - \sum_{i=1}^{k} n_i \bar{x} \\ &= n\bar{x} - \bar{x}\sum_{i=1}^{k} n_i \\ &= n\bar{x} - n\bar{x} \\ &= 0. \end{aligned}$$

Propietate honek geometrikoki $\bar{x}$ datuen grabitate-zentroa dela adierazten du.

3.2 definizioa. Laginean behaturiko balioak era gorakorrean ordenatu ondoren, lagina bi zati berdinetan banatzen duen balioari *mediana* (Me) esaten zaio.

3.3 adibidea. Demagun ordenagailu mota batean lehen matxura gertatu arteko denborak (ordutan) neurtu direla:

322	327	365	371	385	385	393

Egindako $n = 7$ behaketen erdietan, ordenagailuek lehen matxura $Me = 371$ ordu edo lehenago izan dute, eta beste erdietan, $Me = 371$ ordu edo beranduago.

Mediana kalkulatzeko:

- Demagun $x_1, \dots, x_n$ balioak ditugula. Horiei dagokien medianaren balioa kalkulatzeko:

  1. Balioak txikienetik handienera ordenatu.
  2. n bakoitia bada, lagineko $[n/2]+1$-garren balioa izango da mediana. ($[\cdot]$-ek «zati osoa» adierazten du. Adibidez, $[25.7] = 25$).
  3. n bikoitia bada, $[n/2]$-garren eta $[n/2]+1$-garren behaketen erdiko puntua izango da mediana.

  Oharra: aipaturikoa ez da mediana kalkulatzeko modu bakarra.

- Gerta daiteke zenbaitetan lagineko balioak tartean multzokatuta edukitzea, baina ez datu gordinak. Kasu horretan, honela kalkula dezakegu mediana:

  1. Mediana zein tartetan dagoen zehaztu. Horretarako, $n/2$-garren alea zein tartetan dagoen bilatu behar da. Demagun $Me \in [l_i\,,\ l_{i+1})$ dela.
  2. Medianaren balioa tarte horretan proportzionalki kokatu.

$$Me = l_i + \frac{n/2 - N_{i-1}}{n_i} a_i.$$

3.4 adibidea. Itzul gaitezen dagoeneko ikusia dugun datu multzora: *eBay webgunean azken hilabetean Smartphone-produktuak enkantean izan dituen salmenta-prezioak (eurotan).*

250	172	180	419.5	148	152	264	177.5
137.5	175	107	399.5	220	134.9	142	134
224.5	142	132	150	212.5	142	194	132.5
61	125	145.5					

Zein da salmenta-prezioen mediana?

Ebazpena

Medianaren balioa kalkulatzeko, $n = 27$ bakoitia denez, $[n/2] + 1 = 13 + 1$, 14garren balioak emango digu mediana, **lehenik eta behin lagineko balioak ordenatu** ondoren. Orduan, $x_1, \dots, x_n$ balioak ordenatzen baditugu, txikienetik handienera, honela gelditzen zaizkigu:

13 salmenta merkeenak					13 salmenta garestienak	
$x_{(1)} = 61$,	$x_{(2)} = 107$,	$\dots$,	$x_{(13)} = 148$,	$x_{(14)} = 150$,	$x_{(15)} = 152, \dots$,	$x_{(27)} = 419.5$
balio txikiena	2. txikiena	...		↑ **14. alea**		balio handiena

Beraz, $Me = 150$. Hau da, salmenten erdia 150 euro edo merkeago saldu da, eta salmenten beste erdia 150 euro edo garestiago.

3.5 adibidea. Demagun orain 3.4 adibideko balioak tartean ordenatuta topatu ditugula.

$[l_i, l_{i+1})$	n_i	N_i	f_i	F_i	x_i	a_i
$[60, 120)$	2	2	0.074	0.074	90	60
$[120, 180)$	16	18	0.593	0.667	150	60
$[180, 240)$	5	23	0.185	0.852	210	60
$[240, 300)$	2	25	0.074	0.926	270	60
$[300, 360)$	0	25	0	0.926	330	60
$[360, 420)$	2	27	0.074	1	390	60
Guztira	27		1			

Kalkula dezagun medianaren balioa aipatutako eran.

Ebazpena

Lehenik eta behin, mediana zein tartetan kokatzen den erabaki beharko dugu. Horretarako, $n/2 = 13.5$-garren alea zein tartetan kokatuko litzatekeen erabaki behar dugu. Maiztasun metatuen zutabean begiratzen badugu, lehenengo 18 aleak $[120, 180)$ tarteko balioetan edo txikiagoetan kokatzen direla erakusten dugu. Orduan, gu bilatzen ari garen aleak $[120, 180)$ tartean behar luke. Hau da,

$$Me \in [120, 180).$$

Beraz,

$$\begin{aligned} Me &= l_i + \frac{n/2 - N_{i-1}}{n_i} a_i \\ &= 120 + \frac{13.5 - 2}{16} 60 \\ &= 163.125 \end{aligned}$$

Datuak tarte horietan laburbilduta izanik, $Me = 163.1$ izango dugu.

Medianaren zenbait propietate estatistiko:

- Kalkulatzeko erraza da.
- Geometrikoki histogramaren azalera bi zati berdinetan banatzen duen aldagaiaren balioa da.
- Banaketa asimetrikoak dituguneanean, joera zentraleko estatistikorik adierazgarriena da.
- Aurreko propietatearen haritik, ez da muturreko balioekiko sentikorra.



Pentsa ezazu...

Izan bedi lagin hau: $x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 5, x_5 = 5, x_6 = 6, x_7 = 300$.
Kasu honetan, $Me = 5$ da. Gogoratu $\bar{x} = 46.64$ kalkulatu genuela datu horientzat. Bietako zeinek uste duzu adierazten duela hobeto laginaren erdialdea zein balioren inguruan dagoen?

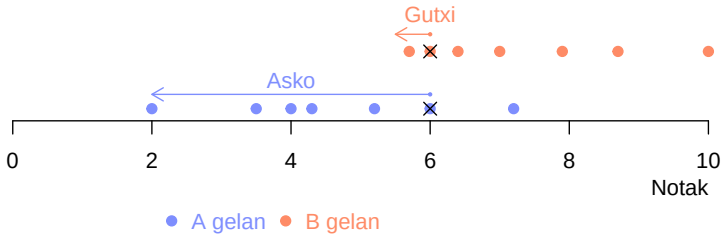
3.2 Posizio estatistikoak

Atal honetan ikusiko ditugun estatistikoek, aldagaiaren behatutako balioak txikienetik handienera ordenatu ondoren, lagineko aleen posizioa eskainiko digute.

3.6 adibidea. Demagun bi ikasgeletako notak ditugula.

A gelako notak	B gelako notak
$2, 3.5, 4, 4.3, 5.2, 6, 7.2$	$5.7, 6, 6.4, 7, 7.9, 8.7, 10$

Begira diezaiogun 6 balioari. 6 balioak A gelako notetan posizio oso altua erakusten du; gelako notarik hoberenetakoa izan da. Hau da, txikiagoak diren nota asko ditu aurrerik. Aldiz, B gelan, 6 balioa posizio baxuetan kokatzen da, gelako notarik txarrenetakoa izan da. Oso balio gutxi ditu aurretik (Ikus 3.2 irudia). Horixe da posizio estatistikoek egiten dutena, aldagaiaren balioak aleen lagineko posizioekin erlazionatu.



3.1 irudia. A gelan notaren 6 balioa posizio altuekin lotzen da baina B gelan, notaren 6 balio bera, posizio baxuekin.

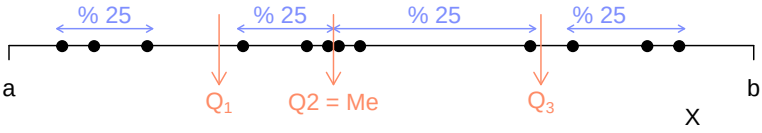
3.3 definizioa. Aleen portzentaje jakin bat ezkerrerantz uzten duen aldagaiaren balioari *koantila* esaten zaio.

Koantilen artean *kuartilak, dezilak* eta *pertzentilak* desberdindu daitezke, jarraian definituko ditugunak.

3.4 definizioa. Lagina 4 zati berdinetan banatzen duten aldagairen balioei *kuartila* esaten zaie. Hiru dira kuartilak:

- Q_1, *lehen kuartilak* laginaren % 25 ezkerrerantz uzten du. Hau da, laginaren laurden batek Q_1 balioa edo txikiagoa du.
- *Bigarren kuartilak* laginaren % 50 ezkerrerantz uzten du. $Q_2 = Me$.
- Q_3, *hirugarren kuartilak* laginaren % 75 ezkerrerantz uzten du.

Ikus kuartilen interpretazio grafikoa 3.2 irudian.



3.2 irudia. Kuartilen definizioa.

3.5 definizioa. Lagina 10 zati berdinetan banatzen duten aldagaiaren balioei *dezila* esaten zaie. Bederatzi dira dezilak, $D_1, \ldots, D_9$:

- D_i, *i. dezilak* laginaren % $(i \times 10)$ ezkerrerantz uzten du, $i = 1, \ldots, 9$. Adibidez, D_1 balioa edo txikiagoa dutenak laginean % 10 dira, aldiz, D_9 edo balio txikiagoa hartzen dutenak % 90 dira.
- $D_5 = Q_2 = Me$.

3.6 definizioa. Lagina 100 zati berdinetan banatzen duten aldagaiaren balioei *pertzentila* esaten zaie. Laurogeita hemeretzi dira pertzentilak, $P_1, \ldots, P_{99}$:

- P_i, *i. pertzentilak* laginaren % i ezkerrerantz uzten du, $i = 1, \ldots, 99$.
- $P_{25} = Q_1$; $\quad P_{50} = D_5 = Q_2 = Me$; $\quad P_{75} = Q_3$.

Koantilen kalkulua medianaren kalkuluaren ideia berari jarraituz egiten da. Ez dago modu bakarra, eta euskarazko wikipedian topa dezakezu horren guztiaren azalpen zabala (http://eu.wikipedia.org/wiki/Koantil).



Sakontzeko

-salmentarekin zerikusia duten 100 enpresek azken urtean izandako irabaziak (milioi eurotan) hauek izan dira:

16.3	24.5	31.7	42.6	48.8	51.2	54.9	55.3	55.3	61.7
62.7	64.4	66.2	66.5	67.7	69.9	70.2	72.6	72.8	73.3
73.7	74.8	75.4	75.8	77.1	78.9	79.0	79.4	80.4	80.6
81.2	82.5	83.1	83.2	83.5	85.1	87.0	87.4	88.1	88.1
88.9	89.3	90.0	90.3	90.7	91.2	91.5	91.5	91.6	91.6
92.0	95.2	97.5	97.7	98.3	98.9	99.1	99.6	101.6	104.7
104.9	105.4	105.5	106.2	106.8	108.9	110.0	110.5	111.0	113.9
114.3	115.7	116.1	116.3	117.7	117.9	120.1	120.3	121.2	122.2
124.2	124.9	125.1	126.3	126.9	128.8	130.0	131.2	133.9	136.8
136.9	140.2	140.4	140.4	140.6	141.6	147.8	149.0	159.9	165.8

1. Kalkula itzazu Q_1, Q_3, P_{10} eta P_{90}. Zer informazio ematen dizute?
2. Baldin eta datu horiek berak $[15, 35), [35, 55), [55, 75), [75, 95), [95, 115), [115, 135), [135, 155), [155, 175)$ tarteetan sailkatuta jasoko bazenitu, nola kalkulatuko zenituzke aurreko atalean eskaturiko estatistikoak?
3. Irudika itzazu datuak grafiko egoki batean, eta koka itzazu bertan Q_1, Q_3, P_{10} eta P_{90}en balioak.

3.3 Sakabanapen estatistikoak

Atal honetan, datuen sakabanapena nolakoa den aztertu eta neurtzeko estatistikoak jorratuko ditugu. Adibide batekin hasiko gara.

3.7 adibidea. Demagun bi ikasgeletako notak ditugula.

A gelako notak	B gelako notak
4.6	1
4.8	1.8
4.9	3
5	5
5	5
5.1	7
5.2	8.2
5.4	9

Joera zentraleko estatistikoei begiratzen badiegu, bi kasuetan notak 5 balioaren inguruan dabiltzala esaten digute ($\bar{x}_A = 5$; $Me_A = 5$ eta $\bar{x}_B = 5$; $Me_B = 5$). Baina hori horrela izanik, bi geletan gertatu dena oso desberdina da. B taldean emaitzak oso sakabanaturik daude, eta A taldean, berriz, oso gutxi. Errealitate horiek neur ditzaketen estatistikoak aztertzea dugu helburu orain.

Irudika itzazu gela bakoitzeko notak zuzen batean, eta oso nabarmen ikusiko duzu nolakoak diren datuak sakabanapenari dagokionez gela bakoitzean.

Asko dira sakababapen estatistikoak, baina denek interpretazio berari jarraitzen diote:

Datuek aldakuntzarik ez dutenean, hau da, balio bakar batean kontzentraturik daudenean, edozein sakabanapen estatistikoren balioa 0 izango da. Beste kasuetan, balio positiboak bakarrik hartzen dituzte. Oro har, zenbat eta sakabanapen handiagoa egon datuetan, orduan eta handiagoa izango da sakabanapen estatistikoaren balioa.

Hauek dira guk ikasiko ditugun sakabanapen estatistikoak:

- Heina.
- Kuartil arteko heina.
- Lagin-bariantza eta kuasibariantza.
- Desbideratze estandarra.
- Aldakuntza-koefizientea.

3.7 definizioa. Izan bedi $x_1, \dots, x_n$ lagina. Laginaren *heina* (R) honela definitzen da:

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

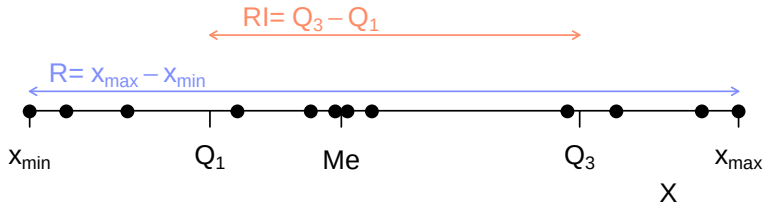
$x_{max} = \max_i\{x_i\}$ eta $x_{min} = \min_i\{x_i\}$ direlarik.

Heina ez da propietate estatistiko hoberenak dituena; izan ere, muturreko balioekiko sentikorra da, eta, oro har, lagina handitu ahala, bere balioa ere handitu egiten da. Heinaren ahulezia hori gainditu eta neurri sendoagoa definituko dugu jarraian.

3.8 definizioa. Izan bedi $x_1, \dots, x_n$ lagina. Laginaren *kuartil arteko heina* (RI) honela definitzen da (Ikus 3.3 irudia):

$$RI = Q_3 - Q_1,$$

Q_1 eta Q_3 lehen eta hirugarren kuartilak dira.



3.3 irudia. Aldagaiaren heinaren eta kuartil arteko heinaren definizioa.

3.9 definizioa. Laginean neurturiko aldagai estatistikoaren *lagin-bariantza* honela definitzen da:

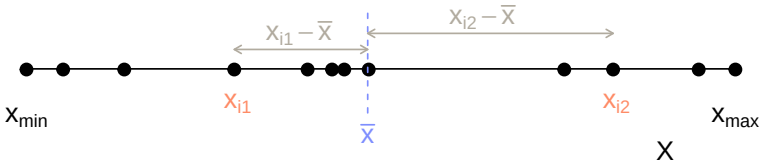
- Datuak elkartuta badaude, hau da, $x_1, \dots, x_k$ behatutako balioen maiztasunak $n_1, \dots, n_k$ badira, hurrenez hurren:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2.$$

- Elkartu gabeko datuak badira, $x_1, \dots, x_n$:

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

Lagin-bariantzak zer neurtzen duen hobeto ulertzeko, ikus 3.4 irudia.



3.4 irudia. Aldagaiaren lagin-bariantzak zer neurtzen duen ulertzeko, batugai bakoitza zer den adierazi da grafikoan.

Askotan oso antzekoa den *kuasibariantza* ere erabiltzen da. Honela definitzen da:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 \quad \text{edo} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2,$$

datuak elkartuta edo banaka izatearen arabera. Kuasibariantzarako $s^2_{n-1} = s^2$ notazioa ere erabiliko dugu. Bi estatistiko horiek oso antzekoak dira; izan ere,

$$s^2 = \frac{n}{n-1} s_n^2.$$

Irakasgaian aurrerago egiten dugunean, hobeto ikusiko ditugu zergatik eta zertarako ditugun bi estatistiko hain antzekoak.

Bariantzaren kalkulua eskuz egiteko, honako identitate hau lagungarri suertatuko zaigu:

- Datuak elkartuta badaude ($x_1, \dots, x_k$ behatutako balioen maiztasunak $n_1, \dots, n_k$ izanik, hurrenez hurren):

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2. \tag{3.1}$$

- Elkartu gabeko datuak badira ($x_1, \dots, x_n$):

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{3.2}$$

Elkartu gabeko datuak direnerako, (3.2) identitatearen frogapena:

$$\begin{aligned} s_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2+\frac{1}{n}\sum_{i=1}^{n}(-2x_i\bar{x})+\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}\bar{x}^2\sum_{i=1}^{n}1 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}^2+\bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{2}x_i^2-\bar{x}^2. \end{aligned}$$

Bariantzaren eta kuasibariantzaren hainbat propietate estatistiko:

- Erabilpen matematikoa erraza denez, gehien erabiltzen diren sakabanapen-neurrien artean daude.
- Muturreko balioekiko sentikorrak dira.
- Beren unitatea aldagaiaren unitatearen karratua da.

Azken propietate horrek estatistikoaren interpretazioa zaildu egiten duenez, arazotxo hori konpontzen duen sakabanapen estatistikoa definituko dugu:

3.10 definizioa. Laginean neurturiko aldagaiaren *desbideratze estandarra* honela definitzen da:

$$s_n = \sqrt{s_n^2}.$$

Oharra: $s = \sqrt{s^2}$-ri ere desbideratze estandarra esaten zaio.

3.8 adibidea. Dagoeneko askotan erabili dugun adibidera itzuliko gara sakabanapen estatistikoen kalkulua nola egin ikusteko.

Inkesta batean, seme-alaba kopurua zein zen galdetu da. Hauek dira lorturiko erantzunak:

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

- Lagin honetan, zein da heina? Eta kuartil arteko heina?
- Kalkulatu lagin-bariantza eta desbideratze estandarra.

Ebazpena

Heinaren kalkulua:

$$\begin{aligned} R &= x_{max} - x_{min} \\ &= 6 - 0 \\ &= 6. \end{aligned}$$

Hau da, lagineko erantzun guztiak 6 seme-alabako tarte batean barneratzen dira. Era berean, $Q_1 = 1$ eta $Q_3 = 2$ izanik, kuartil arteko heina kalkulatuko dugu:

$$\begin{aligned} RI &= Q_3 - Q_1 \\ &= 2 - 1 \\ &= 1. \end{aligned}$$

Hau da, laginaren erdialdeko balioen % 50 seme-alaba bateko tarte batean barneratzen dira.

Lagin-bariantza kalkulatzeko, (3.1) identitatea erabiliko dugu. Horretarako, dagoeneko eraikita genuen maiztasun-taulan beste zutabe bat gehituko dugu:

x_i	n_i	N_i	f_i	F_i	$n_i \times x_i$	$n_i \times x_i^2$
0	4	4	0.1	0.1	0	0
1	17	21	0.425	0.525	17	17
2	10	31	0.250	0.775	20	40
3	6	37	0.150	0.925	18	54
4	2	39	0.050	0.975	8	32
5	0	39	0	0.975	0	0
6	1	40	0.025	1	6	36
Guztira	40		1		69	179

Orduan,

$$\begin{aligned} s_n^2 &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \\ &= \frac{179}{40} - \left(\frac{69}{40}\right)^2 \\ &= 1.50. \end{aligned}$$

Beraz, lagin-bariantza 1.50 seme-alaba^2 da. Hemen bertan ikusten da balio hori interpretatzearen zailtasuna. Desbideratze estandarra, berriz, $s_n = \sqrt{1.50} = 1.22$ seme-alabakoa da.

Jarraian, bariantza kalkulatuko dugu R erabiliz. Erreparatu, kasu honetan, datuak elkartu gabe jaso direla.

```
# Behatutako balioak jaso objektu batean
x <- c(2, 2, 1, ...) # Osatu

# Lagin-tamaina
n <- length(x)
# Emaitza: 40

# Balio guztien batura
batura <- sum(x)
# Emaitza: 69

# Karratuen batura
karratuenbatura <- sum(x^2) # Zer dago x^2 objektuan?
# Emaitza: 179

# Bariantza
karratuenbatura/n - (batura/n)^2
# Emaitza: 1.50
```

Orain arte ikusitako sakabanapen-neurriei neurri absolutuak direla esaten zaie.

Neurri horiek unitate batzuk dituztenez atzetik, oro har ez dute balio aldagai desberdinen sakabanapena konparatzeko, ez eta aldagai berdina lagin desberdinetan neurturik beren arteko sakabanapena konparatzeko ere. Horretarako, dimentsio gabeko honako neurri hau erabil dezakegu:

3.11 definizioa. Laginean neurturiko aldagaiaren *aldakuntza-koefizientea* honela definitzen da:

$$CV = \frac{s_n}{|\bar{x}|}.$$

Dimentsiorik gabeko neurria da, eta portzetajetan adierazten da sarri. Sakabanapen erlatiboa ere esaten zaio.



Sakontzeko

Erabili dugun (3.1) identitatearen frogapena egin ezazu.

3.4 Batezbestekoaren eta desbideratze estandarraren erabilpen bateratua

Desbideratze estandarra batez bestekoarekiko sakabanapen-neurria izanik, datuak batezbestekoarekiko zenbateraino sakabanaturik dauden esaten digu. Hortik arrunta da $(\bar{x} - k \cdot s_n, \bar{x} + k \cdot s_n)$ moduko tarteak eraikitzea. Tarte horien interpretazioa honako hau da:

$(\bar{x} - k \cdot s_n, \bar{x} + k \cdot s_n)$ tartean gutxienez laginaren $\%(1 - 1/k^2) \times 100$ barneratzen da $(k \geq 2)$.

3.9 adibidea. Demagun lagin batean programa baten exekuzio-denborak aztertzen ari garela, eta batez besteko denbora $\bar{x} = 33.2$ segundo eta desbideratze estandarra $s_n = 3.0$ segundo direla.

Zer informazio ematen digute bi estatistiko horiek batera?

Ebazpena

$k = 2$ aukeratuz, alde batetik, $(\bar{x} - k \cdot s_n, \bar{x} + k \cdot s_n) = (27.2, 39.2)$ tartea lortzen dugu. Bestetik, tarte hori $(1 - 1/k^2) = 0.75$ balioari lotzen zaio aurreko interpretazioan. Bi gauzak uztartuz, badakigu gutxienez laginaren % 75 en exekuzio-denborak tarte horretan kokatzen direla. Hau da, gehien jota, laginaren % 25 batek eduki dezake 27.2 segundotik beherako exekuzio-denbora edo 39.2 segundotik gorakoa.

3.4. BATEZBESTEKOAREN ETA DESBIDERATZE ESTANDARRAREN ERABILPEN BATERATUA

Aipatu berri duguna Txebixev-en desberdintza gisa ezagutzen da, eta honela formaliza dezakegu:

Txebixev-en desberdintza. Izan bedi $x_1, \ldots, x_n$ lagina. Finka dezagun $k \geq 2$ balioa eta bana ditzagun lagineko aleak bi multzotan, $A_1 = \{i : |x_i - \bar{x}| \geq k \cdot s_n\}$ eta $A_2 = \{i : |x_i - \bar{x}| < k \cdot s_n\}$. Orduan, A_1-en maiztasun erlatiboa $f(A_1) \leq 1/k^2$ da.

Frogapena. Bariantzaren espresioa aztertuz,

$$\begin{aligned} s_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n}\sum_{i\in A_1}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i\in A_2}(x_i - \bar{x})^2 \\ &\geq \frac{1}{n}\sum_{i\in A_1}(x_i - \bar{x})^2. \end{aligned}$$

Orduan,

$$\begin{aligned} s_n^2 &\geq \frac{1}{n}\sum_{i\in A_1}(x_i - \bar{x})^2 \\ &\geq \frac{1}{n}\sum_{i\in A_1} k^2 s_n^2 \\ &= \frac{1}{n} s_n^2 k^2 n(A_1) \\ &= s_n^2 k^2 f(A_1), \end{aligned}$$

$n(A_1)$ eta $f(A_1)$, A_1-en maiztasun absolutua eta erlatiboa izanik, hurrenez hurren.

Beraz, frogatu nahi zen bezala, $f(A_1) \leq 1/k^2$ eta, baliokideki, $f(A_2) \geq 1 - 1/k^2$.

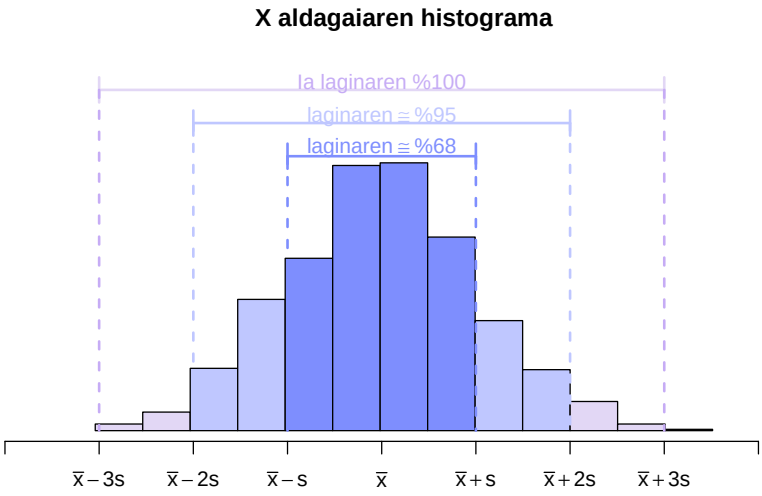
Txebixeven desberdintza edozein egoeratan egiazkoa da, baina zenbait egoeratan ez du oso bornapen zehatza eskaintzen. Bereziki, histogramak kanpai-itxura agertzen duenean, Txebixeven desberdintza baino zehatzagoa da honako erregela enpiriko hau.

Erregela enpirikoa: baldin eta datuek kanpai-itxura agertzen badute:

- Gutxi gorabehera laginaren % 68 batezbestekotik gehienez desbideratze estandar batera dago. Hau da, $(\bar{x}-s, \bar{x}+s)$ tartean laginaren % 68 barneratzen da.
- Gutxi gorabehera laginaren % 95 batezbestekotik gehienez 2 desbideratze estandarrera dago. Hau da, $(\bar{x}-2s, \bar{x}+2s)$ tartean laginaren % 95 barneratzen da.

- Ia lagin osoa batezbestekotik gehienez 3 desbideratze estandarrera dago. Hau da, $(\bar{x} - 3s, \bar{x} + 3s)$ tartean ia lagin osoa barneratzen da.

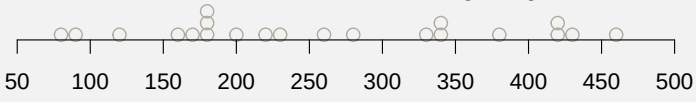
3.5 irudian ikus daiteke erregela enpirikoa grafikoki.



3.5 irudia. Aldagaiaren banaketak kanpai-itxura duenean, $(\bar{x} \mp k \cdot s)$ itxurako tarteen interpretazio geometrikoa.

Sakontzeko

1. Atal honetako helburua aldagai estatistikoa estatistikoekin laburtzen ikastea da. Horretarako, estatistikoak kalkulatzen ikasi duzu, bai eta estatistikoek ematen dituzten emaitzak interpretatzen ere. Liburuxka honen amaieran proposatutako ariketak (3. gaia) egin itzazu helburuak lortzeko.
2. Kutxa-diagrama eraikitzeko behar diren estatistikoak ezagutzen dituzu. Adieraz ezazu puntu-diagraman adierazita dagoen datu multzoa kutxa-diagrama batekin.
   Oharra: nahikoa da estatistikoen kalkulua gutxi gora beherakoa izanda.



3.5 Kutxa-diagrama

2.2 atalean kutxa-diagramak aurkeztu ditugu. Horrelako diagramak sortzeko, ordea, kapitulu honetan ikasitako estatistikoak erabili beharko ditugu.

Zehazki, kutxa-diagrama bat eraikitzeko, honako pauso hauek jarraitu behar ditugu:

- Me, Q_1, Q_3 eta RI kalkulatu.
- Barne-hesiak deitzen diren m_1 eta m_3 puntuak kalkulatu honako era honetara:
$$m_1 = Q_1 - 1.5RI \text{ eta } m_3 = Q_3 + 1.5RI.$$
Puntu horiek outlier-ak edo balio arraroak identifikatzeko erabiliko dira.
- Albo-balioak deitzen diren b_1 eta b_3 balioak kalkulatu, honela:
  - b_1, m_1 balioarengandik hurbilen dagoen aldagaiaren behatutako balioa da, beti ere m_1 baino txikiagoa ez izanik.
  - b_3, m_3 balioarengandik hurbilen dagoen aldagaiaren behatutako balioa da, beti ere m_3 baino handiagoa ez izanik.
- Kanpo-hesiak deitzen diren M_1 eta M_3 puntuak kalkulatu honako era honetan:
$$M_1 = Q_1 - 3RI \text{ eta } M_3 = Q_3 + 3RI.$$
- Erreferentzia-eskala irudikatu, bertikala zein horizontala, eta orain arte kalkulatutako balio guztiak erreferentzia-eskalan kokatu.
- Q_1 eta Q_3 balioak muturtzat dituen kutxa irudikatu, Me parean lerro bat jarririk.
- b_1 eta b_3 albo-balioak «besoen» bidez adierazi, eta horiek kutxarekin lotu.
- Baldin eta barne-hesien eta kanpo-hesien arteko aldagaiaren behatutako baliorik badago, horiek zirkuluen bidez adierazi. Balio horiek *outlier* edo balio arrarotzat hartuko ditugu.
- Baldin eta kanpo-hesien kanpoaldera gelditzen diren aldagaiaren behatutako baliorik balego, asteriskoen bidez adierazi. Balio horiek balio oso arraro edo izugarrizko outlier gisa hartuko ditugu.

4

Estatistika deskribatzailea: bi dimentsiotako banaketak. Korrelazioa eta erregresioa

4.1 Sarrera

Fenomeno bat ezagutzeko, arruntena, aldagai bat baino gehiago batera aztertu beharra izaten da, eta bereziki garrantzitsua suertatzen da aldagaien arteko erlazioak aztertu eta ezagutzea.

4.1 adibidea. Demagun honako galdera hauek erantzun nahi ditugula.

- **G1** Erretzea biriketako minbiziarekin erlazionatuta al dago?
- **G2** Software jakin batek lan bat bukatzeko behar duen denbora eta diskoko sarrera/irteera kopurua erlazionaturik al daude?

  Aurrera jarraitu aurretik, pentsa ezazu zein aldagai estatistiko aztertu behar diren galdera bakoitzean eta zein motatakoak diren.

G1 galderan, $X \equiv$ *«Pertsona erretzailea da»* eta $Y \equiv$ *«Pertsonak biriketako minbizia du»* aldagaiak azaltzen zaizkigu. Kasu honetan, bi aldagaiak bitarrak dira.

G2 galderan, berriz, $X \equiv$ *«Sarrera/irteera kopurua»* eta $Y \equiv$ *«Lana bukatzeko behar duen denbora»* aldagaiak daude. Kasu honetan, biak kuantitatiboak.

Galdera hori erantzuteko, seguraski, softwareak behar duen denbora sarrera/irteera kopuruaren arabera aztertzea egokiagoa izango da, alderantziz baino. Hau da, kasuan kasu, bi aldagaien arteko erlazioa aztertzerakoan zein aldagai zeinen arabera aztertu nahi den erabaki beharko du ikertzaileak. Horrek guztiak honako definizio honetara garamatza.

4.1 definizioa. Demagun X eta Y bi aldagairen arteko erlazioa aztertu nahi dugula. Y-ren balioak X-ren balioen arabera nolakoak diren aztertu nahi badugu, X-ri *aldagai askea* eta Y-ri *menpeko aldagaia* esango diegu. Gainera, $Y \sim X$ moduan adieraziko dugu Y aldagaia X aldagaiaren arabera aztertu nahi dugula.

Horrelako egoerei aurre egiteko bideak X eta Y aldagaien motaren araberakoak dira. Horren arabera eta planteatu ditugun galderetara itzuliz, kasu bakoitzean aztertu behar diren aldagaiak izaera desberdinekoak direnez, galdera horiek erantzuteko jarraitu beharreko bideak ere desberdinak dira.

Apunte hauetan, bai X eta bai Y aldagaiak kuantitatiboak diren egoera nola aztertu besterik ez dugu landuko. Gainera, gai honetan X-ren eta Y-ren arteko erlazioa modu deskribatzailean soilik aztertuko dugu.



Sakontzeko

1. **G1** galderara itzuliz, zein jarriko zenuke aldagai aske gisa eta zein menpeko gisa?
2. Plantea itzazu bi aldagairen arteko erlazioa aztertu beharra dakarten bi galdera. Esan ezazu aldagai bakoitza zein motatakoa den eta kasu bakoitzean zeinek jokatuko duen aldagai aske gisa eta zeinek menpeko aldagai gisa.

4.2 Bi aldagai kuantitatiboren arteko erlazioa. Korrelazioa eta erregresio-zuzena

Bi aldagairen artean egon daitekeen erlazioa aztertzeko, aurreko gaietan bezala, grafikoak eta estatistikoak ditugu. Gaian murgiltzeko adibide bat hartuko dugu.

4.2 adibidea. Gure artean Interneten erabilera asko zabaldu bada ere eta Mendebaldeko herrialdeetan gutxi-gehiago populazioaren erdiak Internet erabiltzen badu ere, hori ez da horrela lurralde guztietan. Ikerketa baten arabera, Saudi Arabian %1ek bakarrik erabiltzen du Internet, eta Nigerian, berriz, %0.1ek. Zein aldagai dago erlazionatuta Interneten erabilerarekin? Honako hauek izan daitezke:

- *GPD*: Barne-produktu gordina, per capita, milaka dolaretan.
- CO_2: Karbono dioxido isuria.
- *Cellular*: Telefono mugikorra duten helduen portzentajea.
- *Fertility*: Emakume helduen batez besteko ume kopurua.
- *Literacy*: Populazioaren alfabetatutako portzentajea.

Demagun $Y \equiv$ *«Internet erabiltzen duen populazioaren portzentajea»* eta $X \equiv$ *«Barne-produktu gordina»* aldagaien arteko erlazioa aztertu nahi dugula, $Y \sim X$ ikuspegiarekin. Horretarako, honako galdera hauek erantzuten saiatu behar dugu.

- Zein grafiko motak erakuts dezake bi aldagaiak erlazionaturik dauden?
- Zein estatistikok neur dezake bi aldagaien arteko lotura?

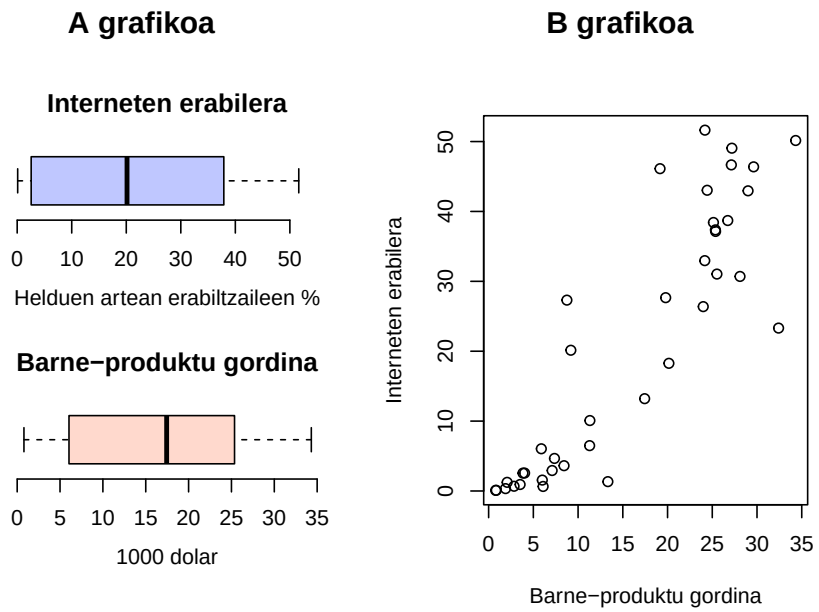
Fitxategia: **human_development.txt**

Fitxategiko datuak kapitulu honen 4.2.2 atalean eta irakasgaiko moodle plataforman topatuko dituzu.

Iturria: *Human Development Report 2003*, United Nations Development Programme, http://hdr.undp.org

Bi aldagai kuantitatiboren arteko erlazioa grafikoki

Gure helburua bi aldagai kuantitatiboren arteko erlazioa aztertzeko grafiko egokia eraikitzea da. Hau da, grafikoa egokia bada, erakusten dizunaren arabera gai izango zara aldagaien artean lotura badagoen ala ez erantzuteko. Aurreko adibidera itzuliz, 4.1 irudian gure helburua bete nahi duten bi grafiko, A eta B, topatuko dituzu.



4.1 irudia. Bi aldagairen (*Barne-produktu gordina* eta *Interneten erabilera*) banaketak adierazteko bi proposamen: A grafikoa ezkerreko zutabean, B grafikoa eskuinekoan.



Pentsa ezazu...

- Zer erakusten dizu A grafikoak? Egokia al da gure helbururako?
- Zer erakusten dizu B grafikoak? Egokia al da gure helbururako?

Galdera horiek ongi erantzun badituzu, badakizu zein grafiko mota den egokia $Y \sim X$ erlazioa aztertzeko, bi aldagaiak kuantitatiboak direnean. Grafiko mota horri *sakabanapen-diagrama* esaten zaio.

Bi aldagai kuantitatiboren arteko erlazioa neurtzeko estatistikoak. Kobariantza eta korrelazioa

Demagun n tamainako lagina dugula eta X eta Y bi aldagai kuantitatibo neurtu direla. Orduan, bildutako datuak honako itxura honetako taula batean antola ditzakegu.

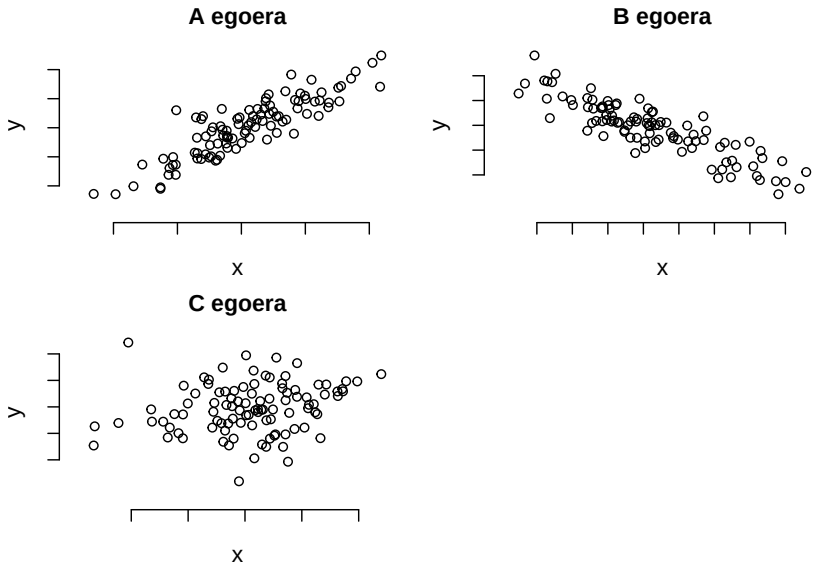
Alea	X	Y
ω_1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
ω_i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
ω_n	x_n	y_n

Laburbilduz, lagina (x_i, y_i), $i = 1, \dots, n$, bikote moduan adieraz dezakegu. Jarraian, 3 egoera desberdini dagozkien grafikoak (4.2 irudia) dituzu, eta, horiei begiratuz, erantzun galdera hauek:

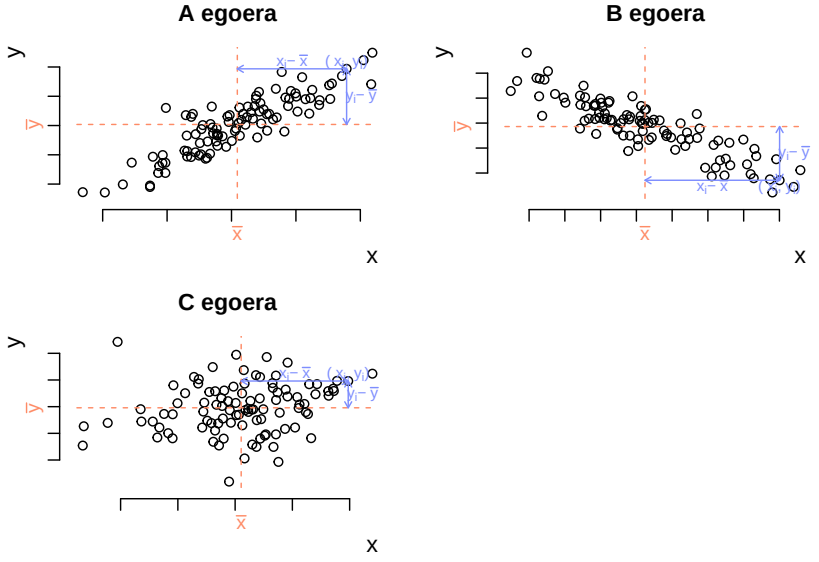
- Zein egoeratan ikusten da X eta Y aldagaien artean ez dagoela loturarik?
- Zein egoeratan ikusten da X eta Y aldagaien artean erlazio zuzena?
- Eta alderantzizkoa?

Ikus dezagun nola eraiki dezakegun erlazio hori edo erlazio eza neur dezakeen estatistikoa. Hasteko, bana ditzagun aleak lau koadrantetan, sakabanapen-diagrama bakoitzean jatorria $(\bar{x}, \bar{y})$ grabitate-zentroan kokatu ondoren. Begira diezaiegun $x_i - \bar{x}$ eta $y_i - \bar{y}$ desbideratzeei, $i = 1, \dots, n$, eta azter dezagun nolakoak diren koadrantez koadrante (ikus 4.3 irudia).

Hartu dezagun $B = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ batukaria, eta saiatu gaitezen bere balioak nolakoak izan daitezkeen ulertzen. Horretarako, kontuan izan nolakoak



4.2 irudia. (X, Y) aldagaien sakabanapen-diagramen hiru egoera desberdin.



4.3 irudia. (X, Y) aldagaien sakabanapen-diagramen hiru egoera desberdin eta $(\bar{x}, \bar{y})$ puntuaren araberako koadranteak.

	Koadranteak			
	I	II	III	IV
$\mathrm{Sgn}(x_i - \bar{x})$	> 0	< 0	< 0	> 0
$\mathrm{Sgn}(y_i - \bar{y})$	> 0	> 0	< 0	< 0
$\mathrm{Sgn}(x_i - \bar{x})(y_i - \bar{y})$	> 0	< 0	> 0	< 0

4.1 taula. Koadranteen zeinuen taula.

diren desbideratzeen ikurrak koadrantez koadrante; ondorioz, nolakoak diren B batukariko elementuen zeinuak koadrantez koadrante (ikus 4.1 taula).



Pentsa ezazu...

Nolako balioak hartuko ditu batukari horrek A, B eta C kasu bakoitzean? Horretarako, lehenik eta behin, aukera ezazu erantzun egokia honako taula honetan:

Koadrantez koadrante,
B batukarian dagoen batugai kopurua

	I	II	III	IV
A kasua	asko?	asko?	asko?	asko?
	gutxi?	gutxi?	gutxi?	gutxi?
B kasua	asko?	asko?	asko?	asko?
	gutxi?	gutxi?	gutxi?	gutxi?
C kasua	asko?	asko?	asko?	asko?
	gutxi?	gutxi?	gutxi?	gutxi?

Ondoren, erabaki ezazu zein den batukariaren ikurra kasu bakoitzean:

- A kasuan, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ _____
- B kasuan, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ _____
- C kasuan, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ _____

Laburbilduz, batukari horren balioarekin jakin dezakegu X eta Y aldagaien arteko erlazio linealik badagoen, eta baldin badago, zuzena ala alderantzizkoa den. Hori guztia kontuan hartuta, jarraian definituko dugun estatistikoa ulertzeko moduan gaude.

4.2 definizioa. Izan bitez $(x_i, y_i), i = 1, \ldots, n$, X eta Y, bi aldagai kuantitatibori dagozkien neurketak. X-ren eta Y-ren arteko *kobariantza* honela definitzen da:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Bai kobariantza eta bai beste estatistiko batzuen kalkuluak errazago egingo dituzu honako identitate hauekin. Batura hauek, oro har, *karratuen batura* izena hartzen dute.

- X-ren karratuen batura:

$$\begin{aligned} SXX &= \sum_i (x_i - \bar{x})^2 \\ &= \sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2 \end{aligned}$$

- Y-ren karratuen batura:

$$\begin{aligned} SYY &= \sum_i (y_i - \bar{y})^2 \\ &= \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2 \end{aligned}$$

- X-ren eta Y-ren karratuen batura:

$$\begin{aligned} SXY &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ &= \sum_i x_i y_i - \frac{1}{n}\sum_i x_i \sum_i y_i \end{aligned}$$

Kobariantzaren hainbat propietate estatistiko:

- Kobariantzak bi aldagai kuantitatiboren arteko erlazio lineala neurtzen du.
- Kobariantza aldagaien unitateen menpeko da.

Azken propietate horrek bere interpretazioa zaildu egiten duenez, bi aldagai kuantitatiboren arteko erlazioa neurtzeko, maiz, honako estatistiko hau erabiltzen da.

4.3 definizioa. Izan bitez $(x_i, y_i), i = 1, \ldots, n$, X eta Y, bi aldagai kuantitatibori dagozkien neurketak. X-ren eta Y-ren arteko *korrelazio-koefizientea* honela definitzen da:

$$r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2\right]^{1/2}}.$$

4.1 propietatea. *Izan bitez X eta Y bi aldagai kuantitatibo eta r_{XY} dagokien korrelazio-koefizientea. Orduan:*

$$-1 \leq r_{XY} \leq 1.$$

Korrelazio-koefizientearen hainbat propietate estatistiko:

- Bi aldagai kuantitatiboren arteko erlazio lineala neurtzen du.
- Interpretazioa:
  - Baldin eta $r \approx 1$, aldagaien artean erlazio zuzena eta sendoa dago.
  - Baldin eta $r \approx -1$, aldagaien artean alderantzizko erlazioa dago eta sendoa da.
  - Baldin eta $r \approx 0$, aldagaien artean ez dago erlazio linealik.

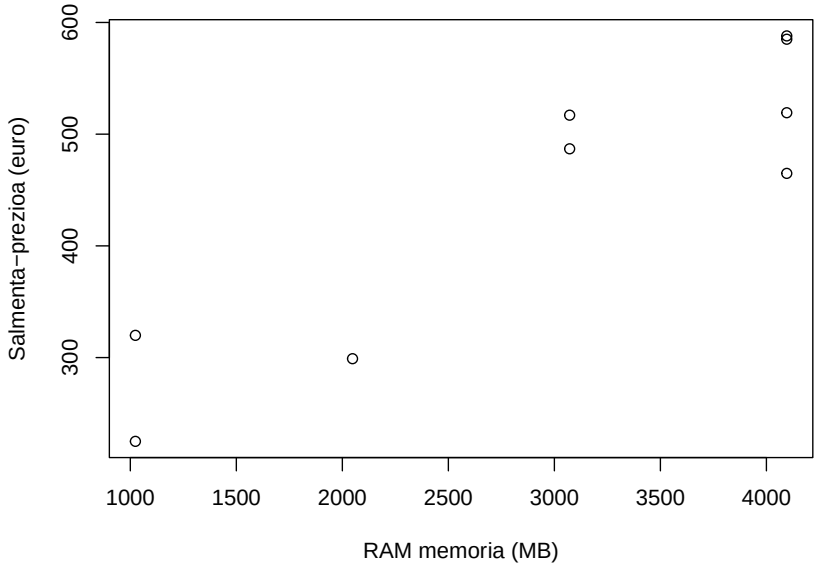
4.3 adibidea. Hona hemen hainbat portatilen salmenta-prezioak (eurotan) eta portatilaren RAM memoria (MB). Portatilen salmenta-prezioa eta RAM memoria dauden aztertzen saiatuko gara. Bildu diren datuak jarraian ditugu.

RAM (x_i)	4096	4096	3072	1024	4096	3072	4096	1024	2048
Prezioa (y_i)	519	464.9	486.9	319.9	585	517	587.96	225	299

- Zein grafiko eraiki dezakezu bi aldagaien artean erlaziorik badagoen aztertzeko?
- Eta zein estatistiko?

Ebazpena

Ikasi dugunez, bi aldagai kuantitatiboren artean erlaziorik badagoen aztertzeko grafiko egokia sakabanapen-diagrama da. Adibide honetan, honako itxura du:



Oro har, bertan ikus dezakegunez, zenbat eta memoria gehiago izan portatilak orduan eta salmenta-prezio altuagoa izan du. Gainera, puntuak zuzen baten inguruan lerrokatuak daudela ere ikus daiteke. Beraz, bi aldagaien artean erlazio lineal zuzena dagoela erakusten du grafikoak.

Erlazio horren sendotasuna neurtzeko, kobariantza eta korrelazio-koefizientea kalkulatuko ditugu. Horretarako, eraiki dezagun honako taula hau:

	y_i	x_i	x_i^2	y_i^2	$x_i\, y_i$
	519.2	4096	16777216	269568.64	2126643.20
	464.9	4096	16777216	216132.01	1904230.40
	486.9	3072	9437184	237071.61	1495756.80
	319.9	1024	1048576	102336.01	327577.60
	585	4096	16777216	342225	2396160
	517	3072	9437184	267289	1588224
	587.96	4096	16777216	345696.96	2408284.16
	225	1024	1048576	50625	230400
	299	2048	4194304	89401	612352
Guztira	4004.86	26624	92274688	1920345.23	13089628.16

Orduan, karratuen baturak hauek dira:

$$\begin{aligned} SXX &= \sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2 \\ &= 92274688 - \frac{1}{9}26624^2 \\ &= 13514980, \\ SYY &= \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2 \\ &= 1920345.23 - \frac{1}{9}4004.86^2 \\ &= 138244.8, \\ SXY &= \sum_i x_i y_i - \frac{1}{n}\sum_i x_i \sum_i y_i \\ &= 13089628.16 - \frac{1}{9}26624 \cdot 4004.86 \\ &= 1242362. \end{aligned}$$

Hemendik, kobariantza eta korrelazio-koefizienteak kalkulatuko ditugu:

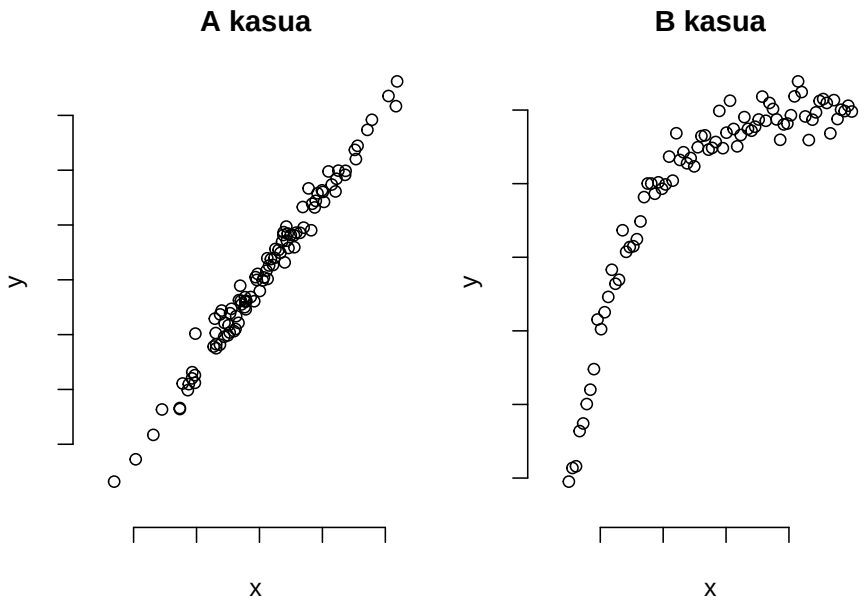
$$\begin{aligned} S_{XY} &= \frac{1}{n-1} SXY \\ &= 155295.3 > 0, \end{aligned}$$

$$\begin{aligned} r_{XY} &= \frac{SXY}{\sqrt{SXX \cdot SYY}} \\ &= 0.9089. \end{aligned}$$

Kobariantza positiboa da eta, ondorioz, baita korrelazio-koefizientea ere. Gainera, $r_{XY} \approx 1$ denez, erlazio zuzena eta sendoa da bi aldagaien artean. Hori guztia bat dator sakabanapen-diagramak erakusten duenarekin.

4.2.1 Erregresio zuzena. Karratu txikienen metodoa

Errepara ezazu X eta Y aldagaien arteko erlazioa nolakoa den 4.4 irudiko bi kasu hauetan.



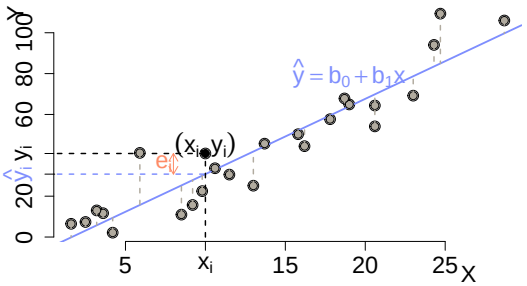
4.4 irudia. (X, Y) aldagaien sakabanapen-diagramen bi kasu desberdin.

Bi kasuetan, X eta Y aldagaien artean erlazio bat antzematen da. A kasuan erlazio hori zuzen batek laburbidu dezake, baina ez B kasuan.

X-ren eta Y-ren arteko erlazioa zuzen batek laburbil dezakeenean, gure helburua hain justu zuzen hori bilatzea da, eta horixe da atal honetan landuko duguna.

Demagun datu multzoan $(x_1, y_1), \ldots, (x_n, y_n)$ balioak behatu ditugula. Gure puntu-hodeia ongien laburbiltzen duen zuzena topatu nahi dugu. Hau da, x_i balioa emanik, $b_0 + b_1 x_i$ zuzenak y_i balioa ahalik eta ongien hurbildu behar du eta horrek (x_i, y_i), $i = 1, \ldots, n$ bikote guztietarako gertatu behar du. Orduan, $\hat{y} = b_0 + b_1 x$ zuzenari X-ren gaineko *Y-ren erregresio-zuzena* esaten zaio.

Erregresio-zuzena topatzea zuzenaren b_1 maldarentzat eta b_0-rentzat balio egokienak bilatzean datza. Horretarako, x_i bakoitzerako behatutako y_i balioak eta erregresio-zuzenak emandako $\hat{y}_i = b_0 + b_1 x_i$ balioak ahalik eta antzekoenak izan behar lukete. Bi balio horien arteko diferentziari *hondarra* esaten zaio, hau da: $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$. Erreparatu, n behaketa ditugunez, n hondar izango ditugula. Oraintxe aipatu ditugun kontzeptuak 4.5 irudian dituzu.



4.5 irudia. Erregresio-zuzena eraikitzeko kontuan eduki beharreko elementuak.

Beraz, gure helburua betetzeko, bila ditzagun b_1 eta b_0 balioak, zeinak erabiliz lortuko ditugun hondarrak ahalik eta txikienak izango baitira. Horretarako, honako funtzio hau minimo egiten duten b_1-en eta b_0-ren balioak topatzea izango da gure helburua:

$$\begin{aligned} G &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

Lehenengo erreparatu beharreko kontua da, G funtzioa b_0 eta b_1 balioen araberako funtzio erreala dela. Hau da,

$$\begin{array}{rccc} G: & \mathbb{R} \times \mathbb{R} & \to & \mathbb{R}^+ \\ & (b_0, b_1) & \mapsto & G = G(b_0, b_1) \end{array}$$

dela. Hori horrela izanik eta Analisi Matematikoa irakasgaian ikasitakoa gogoratuz,



Pentsa ezazu...

- Zer egingo zenuke funtzio erreal batek minimoa zein baliotan erdiesten duen kalkulatzeko?

4.2 propietatea. *Karratu txikienen metodoak eskaintzen duen erregresio-zuzenaren koefizienteak honako hauek dira:*

Zuzenaren malda:

$$\begin{aligned} b_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{SXY}{SXX} \end{aligned}$$

Ordenatu-ardatza mozten duen balioa:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Erregresio-zuzenaren interpretazioa: $\hat{y} = b_0 + b_1 x$ erregresio-zuzena eraikitzen dugunean, $X = x_0$ balio finkorako $\hat{y} \mid_{X=x_0} = b_0 + b_1 x_0$ balioa ematen digu ereduak. Honela, $X = x_0$ ezaugarria duten aleentzat, Y aldagaiaren *batez besteko balioa* aurresaten du ereduak: $\hat{y} \mid_{X=x_0}$.

Batezbesteko horren inguruan Y aldagaiak nolako sakabanapena duen neurtzeko, $y_i - \hat{y}_i = e_i$ hondarrak erabili behar dira.

4.4 definizioa. *Estimazio-errore estandarra* hondarren desbidazio estandarra da, eta honela definitzen da:

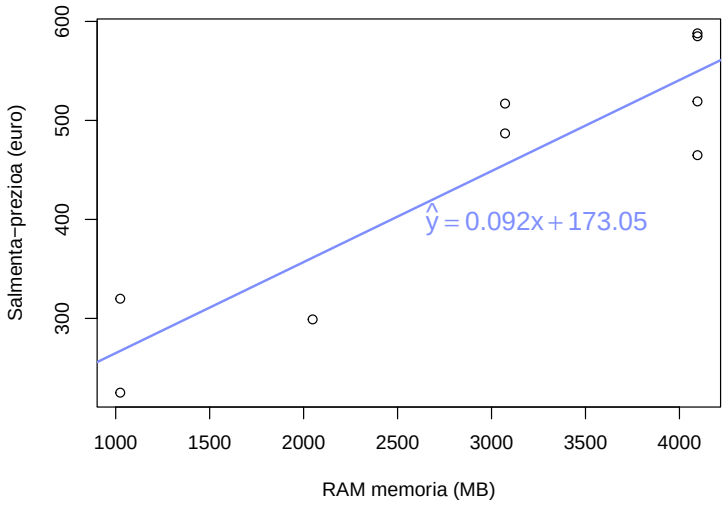
$$s_e = \sqrt{\sum_i (y_i - \hat{y}_i)^2/(n-2)} = \sqrt{RSS/(n-2)}.$$

4.4 adibidea. Erregresio-zuzenaren kalkulua

Aurreko (4.3) adibidean dagoeneko ikusi dugu bi aldagaien artean badagoela erlazio lineal bat. Erlazio lineal hori laburbilduko duen $\hat{y} = b_0 + b_1 x$ erregresio-zuzena kalkulatuko dugu:

$$\begin{cases} b_1 = \frac{SXY}{SXX} = \frac{1272362}{13514980} = 0.092 \\ b_0 = \bar{y} - b_1 \bar{x} = \frac{4004.86}{9} - 0.092 \cdot \frac{26624}{9} = 173.05 \end{cases}$$

Irudi honetan ikus dezakezu erregresio-zuzena puntu-hodeiarekin batera:



Eredu horrek bi aldagaien arteko erlazioa egoki laburbiltzen duenez, eredu hori erabil dezakegu batezbesteko baldintzatuak estimatzeko. Adibidez, $X = 2048$ MB memoria duen portatil batentzat, batez besteko zein prezio espero dugu?

$$\begin{aligned} \hat{y}\,|_{X=2048} &= 0.092 \times 2048 + 173.05 \\ &= 361.5, \end{aligned}$$

Beraz, erregresio-eredu horren arabera, $X = 2048$ MB memoria duen portatil batentzat, batez beste 361.5 euro ordaintzen da.



Sakontzeko

1. Demagun jarraian ikus ditzakezun $n = 4$ behaketa soilik dituzula:

(x_i, y_i)
$(1.0, 1.0)$
$(2.0, 1.4)$
$(1.5, 1.9)$
$(2.5, 2.1)$

Gainera, 3 zuzen proposatu dizkizute:

- Lehenengoa: $b_0 = 0$ eta $b_1 = 1$ koefizienteei dagokiena,

$$\widehat{y}^1 = 0 + 1 \times x.$$

- Bigarrena: $b_0 = 3$ eta $b_1 = -0.75$ koefizienteei dagokiena,

$$\widehat{y}^2 = 3 - 0.75 \times x.$$

- Hirugarrena: $b_0 = 0.62$ eta $b_1 = 0.56$ koefizienteei dagokiena,

$$\widehat{y}^3 = 0.62 + 0.56 \times x.$$

(a) Irudikatu zure datuei dagokien sakabanapen-diagrama. Inongo kalkulurik egin gabe, proposatu zaizkizun 3 zuzenen artean zein uste duzu ez dela egokia inondik inora X eta Y aldagaien arteko erlazioa laburbiltzeko? Arrazoitu erantzuna.

(b) Irudika itzazu 3 zuzenak sakabanapen-diagramaren gainean. Zein uste duzu dela egokiena?

(c) Zuzen bakoitzerako ($Z = 1, 2, 3$) osa ezazu honako itxura honetako taula:

(x_i, y_i)	$\widehat{y}_i^Z$	e_i	e_i^2
$(1.0, 1.0)$			
$(2.0, 1.4)$			
$(1.5, 1.9)$			
$(2.5, 2.1)$			

Taula horietatik, zuzen bakoitzerako kalkula ezazu $G(b_0 = _, b_1 = _) = \sum_{i=1}^{4} e_i^2$ funtzioaren balioa. Orduan, karratu txikienen kriterioaren arabera, zein zuzen da egokiena?

2. Karratu txikienen metodoaren bidez, gauzatu itzazu b_0 eta b_1 balioak lortzeko kalkuluak.

4.2.2 Erregresio-zuzena. Bariantzaren deskonposaketa

Erregresio-ereduarekin menpeko aldagaia helburu jarrita, aldagai horrek erakusten duen aldakortasunari erreparatuko diogu.

Aldakortasuna neurtzeko hainbat estatistiko ikusi ditugu aurreko ataletan, baina, oraingoan beste ikuspegi bat jorratuko dugu.

1. Demagun gure menpeko aldagaiaren, Y-ren, balio hauek behatu ditugula: $y_1, y_2, \dots, y_n$. Gure datu multzoan ez dagoen ale baten Y balioa zein den aurresan beharko bagenu, zer esango genuke?

   Bestelako informaziorik eduki ezean, aurresan dezakegu ale horren Y balioa, gutxi-gehiago, $\bar{y}$ izango dela, eta, hortik, Y-ren aldakortasuna aurresandako balio horretatik dauden desbiderapenek emango digute. Beraz, Y-ren aldakortasuna honela neur dezakegu:

   $$SYY = \sum_i (y_i - \bar{y})^2$$

2. Demagun gure Y menpeko aldagaiaz gain X aldagai askea ere neurtu dugula; kasu honetan, n aleen Y balioez gain X aldagaia ere neurtu dugula: $(x_1, y_1), \dots, (x_n, y_n)$.

   Demagun X eta Y aldagaiak erlazionatuta daudela eta $\hat{y}_i = b_0 + b_1 x_i$ Y-ren erregresio-zuzena kalkulatu dugula. Gure datu multzoan ez dagoen ale baten Y balioa aurresan beharko bagenu, jakinik $X = x_0$ dagokiola, zer esango genuke?

   Kasu honetan, erregresio-zuzenak ematen duen informazioaz baliatuz, gutxi-gehiago, $\hat{y}|x_0 = b_0 + b_1 \cdot x_0$ izango dela esatea arrazonagarria da. Informazio hori erabiliz, Y-ren aldakortasuna honela neur dezakegu:

   $$RSS = \sum_i (y_i - \hat{y}_i)^2$$

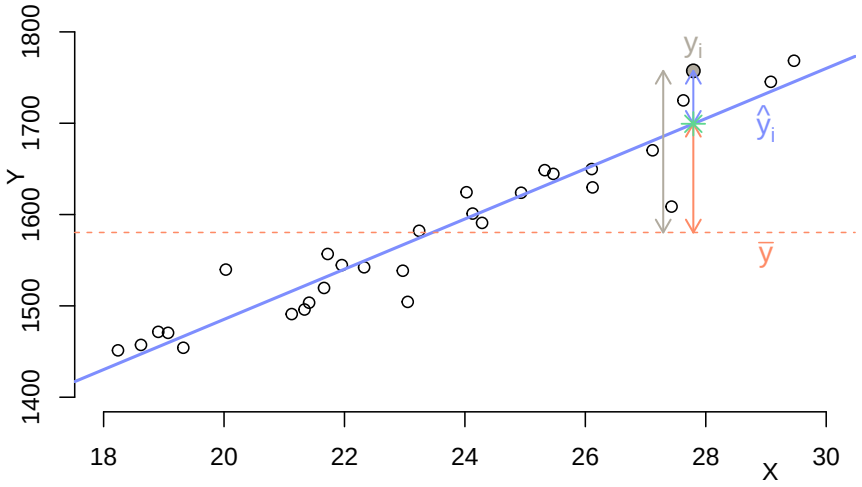
Ikuspegi horiek, *informaziorik ezean*, Y-ren aldakortasuna batetik eta, *X-k ematen digun informazioa kontuan izanik,* Y-ren aldakortasuna bestetik eskaintzen dizkigute. Eta, gainera, bi ikuspegi horiek elkarri lotuta daude.

4.3 propietatea. *Y-ren erregresio-ereduaren arabera, aldakortasunaren honako deskonposaketa hau betetzen da:*

$$SYY = RSS + \sum_i (\hat{y}_i - \bar{y})^2 \tag{4.1}$$

$\sum_i (\hat{y}_i - \bar{y})^2$ gaiari «Erregresio-ereduari esker azaldutako aldakortasuna» esaten zaio, eta $SSreg$ moduan adieraziko dugu. Datu multzoko i ale baterako deskonposaketa hori nolakoa den ikus daiteke 4.6 irudian.

Frogapena. Garatu ezazu zuk. Horretarako, SYY adierazpenean sartu gehitzen eta kentzen diren $\hat{y}_i$ terminoak, eta antolatu gaiak zure helbururako. Behar duzunean, erabili karratu txikienen metodoak eskaintzen duen soluzioa $\hat{y}_i$ terminoan.



4.6 irudia. Erregresio-zuzena kontuan harturik eta i alearen kasurako, Y-ren desbidazioa $\bar{y}$-rekiko $(y_i - \bar{y})$, ereduak aurresandakoarekiko $(y_i - \hat{y}_i)$ eta erregresio-ereduari esker murrizten dugun ziurgabetasuna $(\hat{y}_i - \bar{y})$.

Erregresio-ereduari esker, menpeko aldagaiaren aldakortasuna zenbateraino txikitu daitekeen neurtzeko erabil daiteke (4.1) ekuazioa. Horrela, honako definizio hau dugu:

4.5 definizioa. Erregresio-ereduari dagokion *determinazio-koefizientea* honela definitzen da:

$$R^2 = \frac{SSreg}{SYY}$$

Erregresio-ereduari esker azaltzen den Y-ren aldakortasunaren proportzioa neurtzen du R^2-k.

Propietateak:

- $0 \leq R^2 \leq 1$.

  $RSS = 0$ bada, $R^2 = 1$ puntu-hodeia erregresio-zuzenaren gainean dago.

  $RSS = SYY$ bada, $R^2 = 0$ erregresio-zuzenak ez du Y-ren ezer azaltzen.

- $R^2 = r^2_{XY}$, r_{XY} korrelazio-koefizientea izanik.

Frogapena. Zure esku. Lehen propietatea erraz frogatuko duzu R^2 adierazpeneko terminoen ikurrei erreparatuz. Bigarrenerako, ordezkatu karratu txikienen metodoak eskaintzen dituen balioak b_0 eta b_1 gaietarako, $\hat{y}_i$ terminoan.

4.5 adibidea. Erregresio-ereduaren determinazio-koefizientea

Ikusi berri dugun (4.4) adibidearekin jarraituz, dagoeneko badugu bi aldagaien arteko erlazioa jasotzen duen eredu lineala: $\widehat{y} = 173.05 + 0.092x$.

Eredu hori aintzat harturik, salmenta-prezioaren aldakortasunaren zein portzentaje azaltzen du ordenagailuaren RAM memoriak?

Kalkulatu dugunez, $r_{XY} = 0.9089$ da, eta orduan $R^2 = 0.9089^2 = 0.8261$. Beraz, salmenta-prezioaren % 82.6a RAM memoriari esker azaldu daiteke.

Erregresio-ereduekin bukatzeko, komeni da aipatzea atal honetan ez dugula landu ereduaren egokitasuna. Horretarako, ereduko hondarren azterketa egin behar da, baina irakasgai honetan ez dugu landuko.

human_development.txt fitxategia

human_development.txt[1]

Country	INTERNET	GDP	CO2	CELLULAR	FERTILITY	LITERACY
Algeria	0.65	6.09	3.00	0.30	2.80	58.30
Argentina	10.08	11.32	3.80	19.30	2.40	96.90
Australia	37.14	25.37	18.20	57.40	1.70	100.00
Austria	38.70	26.73	7.60	81.70	1.30	100.00
Belgium	31.04	25.52	10.20	74.70	1.70	100.00
Brazil	4.66	7.36	1.80	16.70	2.20	87.20
Canada	46.66	27.13	14.40	36.20	1.50	100.00
Chile	20.14	9.19	4.20	34.20	2.40	95.70
China	2.57	4.02	2.30	11.00	1.80	78.70
Denmark	42.95	29.00	9.30	74.00	1.80	100.00
Egypt	0.93	3.52	2.00	4.30	3.30	44.80
Finland	43.03	24.43	11.30	80.40	1.70	100.00
France	26.38	23.99	6.10	60.50	1.90	100.00
Germany	37.36	25.35	9.70	68.20	1.40	100.00
Greece	13.21	17.44	8.20	75.10	1.30	96.10
India	0.68	2.84	1.10	0.60	3.00	46.40
Iran	1.56	6.00	4.80	3.20	2.30	70.20
Ireland	23.31	32.41	10.80	77.40	1.90	100.00
Israel	27.66	19.79	10.00	90.70	2.70	93.10
Japan	38.42	25.13	9.10	58.80	1.30	100.00
Malaysia	27.31	8.75	5.40	31.40	2.90	84.00
Mexico	3.62	8.43	3.90	21.70	2.50	89.50
Netherlands	49.05	27.19	8.50	76.70	1.70	100.00
New Zealand	46.12	19.16	8.10	59.90	2.00	100.00
Nigeria	0.10	0.85	0.30	0.30	5.40	57.70
Norway	46.38	29.62	8.70	81.50	1.80	100.00
Pakistan	0.34	1.89	0.70	0.60	5.10	28.80
Philippines	2.56	3.84	1.00	15.00	3.20	95.00
Russia	2.93	7.10	9.80	5.30	1.10	99.40
Saudi Arabia	1.34	13.33	11.70	11.30	4.50	68.20
South Africa	6.49	11.29	7.90	24.20	2.60	85.00
Spain	18.27	20.15	6.80	73.40	1.20	96.90
Sweden	51.63	24.18	5.30	79.00	1.60	100.00
Switzerland	30.70	28.10	5.70	72.80	1.40	100.00
Turkey	6.04	5.89	3.10	29.50	2.40	77.20
United Kingdom	32.96	24.16	9.20	77.00	1.60	100.00
United States	50.15	34.32	19.70	45.10	2.10	100.00
Vietnam	1.24	2.07	0.60	1.50	2.30	90.90
Yemen	0.09	0.79	1.10	0.80	7.00	26.90

[1]Human Development Report 2003, United Nations Development Programme, http://hdr.undp.org

5

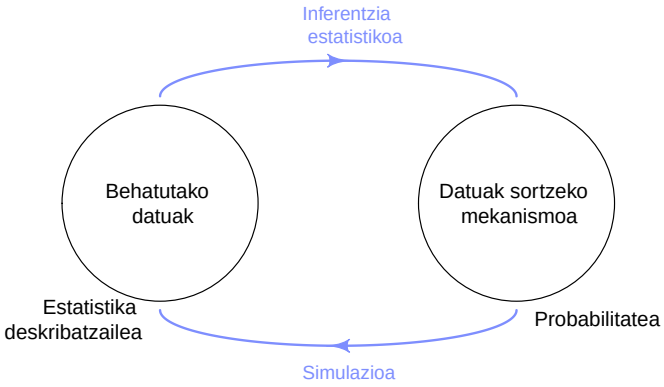
Probabilitate-kalkulua: oinarrizko probabilitate-teoria

5.1 Sarrera

Probabilitateaz hitz egiteko *zorizko esperimentu* baten testuinguruan kokatu behar gara, hau da, esku artean dugun esperimentuak emaitza bat baino gehiago izan ditzakeenean eta horietako zein gertatuko den ezin dugunean aurresan. Oso adibide sinplea eta ulerterraza da *dado bat botatzean* dugun esperimentua. Izan ere, kasu horretan 6 dira emaitza posibleak. baina ezin dugu aurresan horietako zein gertatuko den. Beraz, dado bat botatzea zorizko esperimentua dela esango dugu. Egoera horretan, ezingo dugu aurresan zein izango den esperimentuaren emaitza, baina bai esperimentu horretan dauden erregulartasunak aztertu. Hau da, ezin dugu aurresan, adibidez, «1»a aterako den baina badakigu, adibidez, oro har «1»a eta «6»a neurri berean azalduko direla. Hori izango da «Probabilitatearen» lana, zorizko esperimentuan dauden erregulartasunak aztertzea.

Ikasturte hasieran dagoeneko 5.1 irudiko eskema azaldu genuen. Zorizko esperimentu baten testuingurua hartuz, aipaturiko irudiko *behatutako datuak* egoera esperimentua hainbat aldiz probatu **ondoren** ikusitako balioek osatuko luteke. Aldiz, *datuak sortzeko mekanismoa* egoera esperimentua probatu **aurretik** egin dezakegun azterketari dagokio, eta horretaz arduratuko gara gai honetan.

5.1 definizioa. Zorizko esperimentuaren emaitza posible guztien multzoari *lagin-espazioa* esango diogu, eta Ω moduan adieraziko dugu.



5.1 irudia. Estatistika eta Probabilitatea kokatzeko diagrama.

5.1 adibidea. Ondorengo zorizko esperimentu bakoitzari dagokion lagin-espazioa kalkulatuko dugu.

- Esperimentua: «Dado bat bota»

$$\Omega = \{1, 2, 3, 4, 5, 6\}\,.$$

- Esperimentua: «Txanpon bat bi aldiz bota»

  Txanponaren alde bat A izanik eta bestea B,

$$\Omega = \{(A, A), (A, B), (B, A), (B, B)\}\,.$$

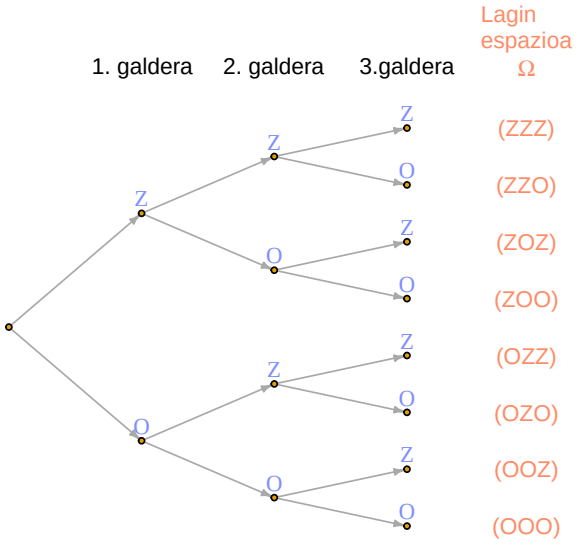
- Esperimentua: «Irakasleak 3 galderako proba bat prestatu du. Galdera bakoitzaren erantzuna *zuzena* (Z) ala *okerra* (O) soilik izan daiteke. Ikasle batek lehen bi galderak zuzen eta azkena oker erantzun baditu, ZZO moduan adieraz dezakegu. Burututako proba bat zoriz aukeratu, eta nolako erantzunak ikus daitezke egoera honetan?»

$$\Omega = \{ZZZ, ZZO, ZOZ, ZOO, OZZ, OZO, OOZ, OOO\}\,.$$

  Jaso daitezkeen erantzun posible guztien multzoa da lagin-espazioa, eta zenbait egoeratan zuhaitz-diagrama bat eraikitzeak erantzun posible guztiak hobeki ikusten lagundu dezake (Ikus 5.2 irudia).

5.2 Gertaerak

Izan bedi Ω zorizko esperimentu bati loturiko lagin-espazioa.



5.2 irudia. Zuhaitz-diagrama, «Irakasleak 3 galderatako proba bat prestatu du» esperimentuko adibiderako.

Ω-ko elementuei *oinarrizko gertaera* esaten zaie. Oinarrizko hainbat gertaeraz osaturiko multzoari *gertaera konposatua* esaten zaio.

5.2 adibidea. Aurreko adibideko (5.1) zorizko esperimentuekin jarraituz.

- Esperimentua: «Dado bat bota»

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

  1. 6a ateratzeari begiratzen badiogu, honela idatz dezakegu:

     $A \equiv$ *«6a ateratzea»*, $A = \{6\}$ eta oinarrizko gertaera da.

  2. Era berean, $B \equiv$ *«bikoitia ateratzea»*, $B = \{2, 4, 6\}$ eta gertaera konposatua da.

- Esperimentua: «Irakasleak prestatutako 3 galderatako proban ikasle batena zoriz aukeratu».

  1. $A \equiv$ *«Ikasleak bi galdera zuzen erantzun ditu»*

     $A = \{ZZO, ZOZ, OZZ\}$

  2. $B \equiv$ *«Ikasleak lehen galdera zuzen erantzun du»*

     $B = \{ZZZ, ZZO, ZOZ, ZOO\}$

Hemendik aurrera, lagin-espazio finituen kasura mugatuko gara ($\sharp\Omega < \infty$).

5.2 definizioa. Ω-ren parteen multzoak, $\mathcal{P}(\Omega)$ moduan adierazitakoak, *gertaera guztien multzoa* adierazten du. Horien artean badira bi gertaera berezi:

- Ω, gertaera *segurua*.
- $\emptyset$, *ezinezko* gertaera.



Pentsa ezazu...

Dado bat botatzeari dagokion lagin-espazioa $\Omega = \{1, \dots, 6\}$ da. Zein da bere parteen multzoa? Zenbat elementu ditu?

5.2.1 Gertaeren arteko erlazioak

Izan bedi Ω lagin-espazioa.

5.3 definizioa.

- A-ren eta B-ren arteko *bildura*,

$$A \cup B = \{x \in \Omega | x \in A \text{ edo } x \in B\}.$$

- A-ren eta B-ren arteko *ebakidura*,

$$A \cap B = \{x \in \Omega | x \in A \text{ eta } x \in B\}.$$

- A-ren *osagarria*,

$$\bar{A} = \{x \in \Omega | x \notin A\}.$$

- A-ren eta B-ren arteko *kendura*,

$$A - B = \{x \in A | x \notin B\}.$$

  Beraz, $A - B = A \cap \bar{B}$.

Jarraian, gertaerak eta beren arteko erlazioen hainbat propietate aipatuko ditugu:

- **De Morgan-en legeak**. Izan bitez $A, B \in \mathcal{P}(\Omega)$.
  1. $\overline{A \cup B} = \bar{A} \cap \bar{B}$
  2. $\overline{A \cap B} = \bar{A} \cup \bar{B}$
  3. $A_1, \dots, A_k$ gertaeretara orokortuz,

$$\overline{\left(\bigcup_{i=1}^{k} A_i\right)} = \bigcap_{i=1}^{k} \overline{A_i} \quad \text{eta} \quad \overline{\left(\bigcap_{i=1}^{k} A_i\right)} = \bigcup_{i=1}^{k} \overline{A_i}\,.$$

- **Banatze-propietateak**. Izan bitez A, B eta $C \in \mathcal{P}(\Omega)$.
  1. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  2. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

5.4 definizioa. Izan bitez A eta B gertaerak. Baldin eta $A \cap B = \emptyset$ bada, A eta B gertaerak *bateraezinak* direla esaten da.



Sakontzeko

Osa itzazu jarraian dituzunak:

1. $A \cup \emptyset =$
2. $A \cap \Omega =$
3. $A \cap \emptyset =$
4. $A \cup \Omega =$
5. $A \cup \bar{A} =$
6. $A \cap \bar{A} =$

5.3 Probabilitate-funtzioa

5.5 definizioa. Izan bedi Ω lagin-espazio finitua E zorizko esperimentuari lotua, eta $\mathcal{P}(\Omega)$ bere parteen multzoa. Izan bedi P funtzioa

$$\begin{array}{rcl} P : \mathcal{P}(\Omega) & \to & [0,1] \\ A & \hookrightarrow & P(A) \end{array}$$

non honako axiomak hauek betetzen baitira:

i. $P(\Omega) = 1$.
ii. $\forall A \in \mathcal{P}(\Omega) \quad P(A) \geq 0$.
iii. $\forall A, B \in \mathcal{P}(\Omega)$ non $A \cap B = \emptyset$ baita,

$$P(A \cup B) = P(A) + P(B).$$

P funtzioak hiru axioma horiek betetzen dituenean *probabilitate-funtzioa* dela esaten da, eta $(E, \mathcal{P}(\Omega), P)$ hirukoteak *probabilitate-espazioa* osatzen dutela.

Axioma horiek honako hau adierazten dute:

i. Gertaera segurua lortzeko ziurtasuna.
ii. Probabilitatea, «gertatzeko aukeraren» neurri gisa ulertuz, negatiboa izateak ez du inolako zentzurik.
iii. Bi gertaera batera lortzea ezinezkoa denean, bi gertaeretakoren bat gertatzeko aukerak dira, bata gertatzeko aukerak gehi bestea gertatzeko aukerak.



Pentsa ezazu...

Axioma horiei begiratuz, zure probabilitateari buruzko ideia intuitiboa bermatzen dutela esango zenuke?

5.1 propietatea. *Izan bitez Ω zorizko esperimentuari loturiko lagin-espazioa eta P probabilitate-funtzioa.*

1. $P(\emptyset) = 0$
2. $\forall A \in \mathcal{P}(\Omega) \quad P(\bar{A}) = 1 - P(A)$
3. $\forall A, B \in \mathcal{P}(\Omega) \; : \; A \subset B \quad P(A) \leq P(B)$ *eta* $P(B - A) = P(B) - P(A)$
4. $\forall A, B \in \mathcal{P}(\Omega) \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   Oharra: *formula hori gertaera kopuru handiagoetarako orokortu daiteke. Adibidez,* $\forall A, B, C \in \mathcal{P}(\Omega)$

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad -P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad +P(A \cap B \cap C) \end{aligned}$$

Frogapena. Propietate horiek frogatzeko, probabilitate-funtzioaren definizioko hiru axiometan oinarritu behar gara.

1. $P(\emptyset) = 0$ frogatu behar dugu.

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \emptyset)\,, \quad \Omega \text{ eta } \emptyset \text{ bateraezinak direnez,} \\ P(\Omega) &\overset{iii.}{=} P(\Omega) + P(\emptyset) \Rightarrow P(\emptyset) = 0_{\square} \end{aligned}$$

2. $P(\bar{A}) = 1 - P(A)$ frogatu behar dugu.
   - $\Omega = A \cup \bar{A}$ denez,

$$\begin{aligned} P(\Omega) &= P(A \cup \bar{A}) \\ 1 &\overset{i.}{=} P(A \cup \bar{A}) \text{ dugu.} \end{aligned}$$

   - A eta $\bar{A} = \emptyset$ bateraezinak direnez, iii. axioma aplikatu dezakegu, eta hortik:

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}).$$

Bi bide horiek uztartuz,

$$1 = P(A) + P(\bar{A}) \Rightarrow P(\bar{A}) = 1 - P(A)_{\square}.$$

3. Bi atal ditugu frogatzeko: batetik, $P(A) \leq P(B)$ eta, bestetik, $P(B - A) = P(B) - P(A)$, beti-ere $A \subset B$ denean.

   Kasu honetan, $B = A \cup (B \cap \bar{A})$ dugu eta, $A \cap (B \cap \bar{A}) = \emptyset$, bateraezinak dira. Honela,

$$\begin{aligned} P(B) &= P(A \cup (B \cap \bar{A})) \\ &\overset{iii.}{=} P(A) + \underbrace{P(B \cap \bar{A})}_{\geq 0} \end{aligned} \tag{5.1}$$

   (5.1) berdintza hartuz, batetik, $P(A) \leq P(B)$ dugu. Bestetik, $B - A = B \cap \bar{A}$ dela gogoratuz,

$$P(B - A) = P(B) - P(A)_{\square}$$

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ frogatu behar dugu.

   - $A \cup B = A \cup (B \cap \bar{A})$ gertaera bateraezinen bildura gisa idatziko dugu. Honela,

$$\begin{aligned} P(A \cup B) &= P(A \cup (B \cap \bar{A})) \\ P(A \cup B) &\overset{iii.}{=} P(A) + P(B \cap \bar{A}). \end{aligned} \tag{5.2}$$

   - $B = (B \cap A) \cup (B \cap \bar{A})$ ere gertaera bateraezinen bildura gisa idatziko dugu. Hemendik,

$$\begin{aligned} P(B) &= P\left(B \cap A) \cup (B \cap \bar{A})\right) \\ P(B) &\overset{iii.}{=} P(A \cap B) + P(B \cap \bar{A}). \end{aligned} \tag{5.3}$$

   Orain (5.2) eta (5.3) berdintzak uztartuz, nahi genuena frogatuko dugu:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)_{\square}$$

5.3 adibidea. Izan bitez A eta B bi gertaera non, $P(A) = 0.2$, $P(B) = 0.4$ eta $P(A \cup B) = 0.5$ baitira. Kalkula ditzagun $\bar{A}$, $A \cap B$, $\bar{A} \cap \bar{B}$ eta $\bar{A} \cup \bar{B}$ gertaeren probabilitateak.

Ebazpena

- $P(\bar{A})$ kalkulatuko dugu:

$$\begin{aligned} P(\bar{A}) &= 1 - P(A) \\ &= 1 - 0.2 \\ &= 0.8. \end{aligned}$$

- $P(A \cap B)$ kalkulatzeko, 5.1 propietateko 4. emaitzara joko dugu. Bertatik,

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= 0.2 + 0.4 - 0.5 \\ &= 0.1. \end{aligned}$$

- De Morganen emaitza erabiliz $\bar{A} \cap \bar{B} = \overline{A \cup B}$ dugu. Hemendik, berriro 5.1 propietatera joz:

$$\begin{aligned} P(\overline{A \cup B}) &= 1 - P(A \cup B) \\ &= 1 - 0.5 \\ &= 0.5. \end{aligned}$$

- Modu berean:

$$\begin{aligned} P(\bar{A} \cup \bar{B}) &= P(\overline{A \cap B}) \\ &= 1 - P(A \cap B) \\ &= 1 - 0.1 \\ &= 0.9. \end{aligned}$$

5.4 Nola finkatu probabilitateen balioak. Zorizko eredu klasikoa

Izan bedi Ω zorizko esperimentu bati dagokion lagin-espazioa. Gertaera bakoitzari dagokion probabilitatea ezagutzeko hainbat teknika erabili daitezke (maiztasun erlatiboen bitartezko estimazioa, adibidez), baina, kasu honetan, *zorizko eredu klasikoa* edo *Laplaceren eredua* aztertuko dugu.

Horretarako, har dezagun $\Omega = \{e_1, \ldots, e_N\}$ lagin-espazio finitua, oinarriko gertaera guztiek probabilitate berdinak dituztelarik, hau da,

$$P(\{e_i\}) = p \quad i = 1, \ldots N.$$

Orduan,

1. $p = 1/N$
2. $\forall A \in \mathcal{P}(\Omega)$,

$$P(A) = \frac{\#A}{N}.$$

Askotan, $\#A$-ri *aldeko kasuak* esaten zaio.

Ikus dezagun honek horrela behar duela. Izan ere, batetik $P(\Omega) = 1$ da; bestetik, $\Omega = \bigcup_{i=1}^{N} \{e_i\}$ dugu. Gainera, oinarrizko gertaerak binaka bateraezinak direnez, probabilitate-funtzioaren definizioko iii. axioma erabil dezakegu.

Laburbilduz,

$$\begin{aligned} P(\Omega) &= P(\bigcup_{i=1}^{N} \{e_i\}) \\ &\overset{iii)}{=} \sum_{i=1}^{N} P(\{e_i\}) \\ &= Np = 1 \end{aligned}$$

Eta hortik, $p = 1/N$ dela ikusiko dugu.

5.4 adibidea. Bi dado botako ditugu. Zein da atetarako emaitzen batura 7 izateko probabilitatea?

Ebazpena

$A \equiv$ *«Dadoen batura 7 da»* $\longrightarrow P(A) = ?$

Lehenik eta behin, esperimentu honi dagokion lagin-espazioa zein den ikusiko dugu. Horretarako:

Lehen dadoa	Bigarren dadoa	Lagin-espazioa
1	1	(1, 1)
	⋮	⋮
	6	(1, 6)
⋮		⋮
6	1	(6, 1)
	⋮	⋮
	6	(6, 6)

Gainera, $A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$.

Honela, $\#\Omega = 6^2 = 36$ eta $\#A = 6$. Beraz,

$$P(A) = \frac{6}{36} = \frac{1}{6}.$$

Oharra. Azken adibide horretan Ω eta A-ko elementu denak izendatu baditugu ere, probabilitatea kalkulatzeko benetan behar duguna multzo horiek dituzten elementu kopurua da. Lan horretarako dugu, hain zuzen ere, *konbinatoria*. Seguru ongi etorriko zaizula «Matematika Diskretua» irakasgaian ikasi zenuena errepasatzea.



Sakontzeko

Dado bat hiru aldiz botako dugu, eta hiru emaitzen baturari erreparatuko diogu.

1. Zein da lortu daitekeen batura txikiena? Eta handiena? Zein dira gertaera horiei dagozkien probabilitateak?
2. Zein da batura 9 izateko probabilitatea? Eta batura 10 izatekoa?

5.5 Probabilitate baldintzatua

Demagun 100 ordenagailu ditugula eta horietako bat zoriz aukeratuko dugula. Ordenagailuen artean 30ek Linuxa dute, eta gainontzekoek Windowsa. Linuxa dutenen artean, 25ek LibreOffice-a dute. Windowsa dutenen artean, berriz, 30ek dute LibreOffice-a.

Zoriz ordenagailua aukeratu, eta jar ditzagun honako gertaera hauek:

- $L \equiv$ *«Linuxa du»*
- $LO \equiv$ *«LibreOffice-a du»*

Batetik, suertatutako ordenagailuak LibreOfficea edukitzeko probabilitatea honako hau da:

$$P(LO) = \frac{25 + 30}{100} = 0.55.$$

Baina suertatuko ordenagailuak Linuxa jarrita duela baldin badakigu, informazio hori kontuan izanda, ordenagailuak LibreOfficea edukitzeko probabilitatea hau da:

$$P(LO \,|\, L) = \frac{25}{30} = 0.833.$$

Hau da, L gertaera izan dela jakiteak, LO gertaeraren probabilitatea informazio hori txertatuta kalkulatzea baimentzen digu.

5.6 definizioa. Izan bitez zorizko esperimentua, Ω lagin-espazio finitua eta P bertan definituriko probabilitate-funtzioa. Hartu dezagun $A \in \mathcal{P}(\Omega)$ gertaera finkoa, $P(A) > 0$ delarik. Orduan, beste edozein B gertaeraren A-rekiko baldintzatutako probabilitatea $P(B|A)$ moduan adierazten da, eta honela definitzen da:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Lehentxeago, 5.5 atalaren hasieran erabili dugun adibidera bueltatuz, ikus dezagun probabilitate baldintzatuaren definizioa erabiliz nola berreskuratzen dugun kalkulatuta genuen probabilitate bera:

$$\begin{aligned} P(LO|L) &\overset{def.}{=} \frac{P(LO \cap L)}{P(L)} \\ &= \frac{25/100}{30/100} \\ &= 0.833. \end{aligned}$$

5.2 propietatea. *Izan bedi $(E, \mathcal{P}(\Omega), P)$ probabilitate-espazioa. Finka dezagun $A \in \mathcal{P}(\Omega)$ gertaera, $P(A) > 0$ delarik. Eraiki dezagun honako funtzio hau:*

$$\begin{aligned} P(\cdot\,|A) : \mathcal{P}(\Omega) &\to [0,1] \\ B &\hookrightarrow P(B\,|A) \end{aligned}$$

Orduan, $P(\cdot\,|A)$ funtzioa ere probabilitate-funtzioa da.

Frogapena. Egin ezazu zeuk. Pentsatu: zer egin beharko duzu probabilitate-funtzioa dela frogatzeko? Hori da! Probabilitate-funtzioaren definizioa ematen duten hiru axiomak betetzen direla frogatu behar duzu.



Pentsa ezazu...

Honako berdintzak hauek egiazkoak al dira? Zergatik?

- $P(\bar{B}|A) = 1 - P(B|A)$
- $P(B \cup C|A) = P(B|A) + P(C|A) - P(B \cap C|A)$
- $P(\bar{B}|A) = 1 - P(B|\bar{A})$

Probabilitate konposatuaren teorema. Izan bitez $A_1, \ldots, A_k$ gertaerak, non $P\left(\bigcap_{i=1}^{k} A_i\right) > 0$ baita. Orduan,

$$P\left(\bigcap_{i=1}^{k} A_i\right) = P(A_1) \times P(A_2|A_1) \times \ldots \times P\left(A_k|\bigcap_{i=1}^{k-1} A_i\right).$$

Frogapena. k-ren gaineko indukzioz frogatuko dugu.

1. $k = 2$ denean, egiazkoa da teoremak dioena?

   Probabilitate baldintzatuaren definizioaren arabera, $P(A_2|A_1) = P(A_2 \cap A_1)/P(A_1)$ da. Hortik berehalakoa da frogatu nahi genuen $P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$.

2. Indukzio-hipotesiaren (i.h.) arabera, egiazkoa da teoremak dioena k-rako.

   Egiazkoa da, orduan, $(k+1)$-erako?

$$\begin{aligned} P\left(\bigcap_{i=1}^{k+1} A_i\right) &= P\left((\bigcap_{i=1}^{k} A_i) \cap A_{k+1}\right) \\ &\overset{*}{=} P\left(\bigcap_{i=1}^{k} A_i\right) \times P\left(A_{k+1}|\bigcap_{i=1}^{k} A_i\right) \\ &\overset{i.h.}{=} P(A_1) \times P(A_2|A_1) \times \ldots \times P\left(A_k|\bigcap_{i=1}^{k-1} A_i\right) \times P\left(A_{k+1}|\bigcap_{i=1}^{k} A_i\right) \end{aligned}$$ □

5.5 adibidea. Bideokamerak ekoizten dituen enpresa batek bi eredu berri atera ditu: M_1 arrunta eta M_2 aurreratua. Salmenten % 40 M_1 eredukoa izan da, eta M_1 eredua erosi dutenen artean % 30-ek garantia zabalagoa kontratatu (G) dute. Gainera, garantia zabalago hori aukeratu dutenen artean bideokamerak arazoren bat izan eta ez konpontzeko (A) probabilitatea 0.001 da. Zein da pertsona batek $M1$ eredua erosi, garantia zabalagoa kontratatu eta arazoak ez konpontzeko probabilitatea?

Ebazpena

Ariketaren enuntziatuak ezartzen dituen gertaera eta notazioa jarraituz, $P(M_1 \cap G \cap A)$ kalkulatu behar dugu.

Enuntziatuak honako probabilitate hauek eman dizkigu:

$$P(M_1) = 0.4, \quad P(G|M_1) = 0.3, \quad P(A|G \cap M_1) = 0.001.$$

Beraz, Probabilitate konposatuaren teorema (PKT) aplikatuz:

$$\begin{aligned} P(M1 \cap G \cap A) &\overset{PKT}{=} P(M_1)P(G|M_1)P(A|G \cap M_1) \\ &= 0.4 \cdot 0.3 \cdot 0.001 \\ &= 1.2 \cdot 10^{-4}. \end{aligned}$$

5.6 Bayes-en teorema

5.7 definizioa. Izan bitez $A_1, \dots, A_k$ gertaerak, non honako hauek betetzen baitira:

- Gertaerak binaka batera-ezinak dira:

$$A_i \cap A_j = \emptyset\,, \quad i \neq j$$

- Gertaerek Ω osoa estaltzen dute:

$$\bigcup_{i=1}^{k} A_i = \Omega$$

Orduan, $\{A_1, \dots A_k\}$-ek *gertaera sistema osoa* (g.s.o.) osatzen dutela esaten da.

5.6 adibidea. Esperimentua: «Dado bat bota»

- $A_1 \equiv$ *«Bikoitia»*
- $A_2 \equiv$ *«Bakoitia»*

$\{A_1, A_2\}$ -k g.s.o osatzen dute; izan ere, $A_1 \cap A_2 = \emptyset$ eta $A_1 \cup A_2 = \Omega$.

Erreparatu $A_2 = \bar{A_1}$ dela. $\{A, \bar{A}\}$-k beti osatzen dute g.s.o.

Probabilitate osoaren teorema. Izan bedi $\{A_1, \dots, A_k\}$ g.s.o.

$P(A_i) > 0$ izanik, $i = 1, \dots, k$. Orduan, $\forall B \in \mathcal{P}(\Omega)$ harturik,

$$\begin{aligned} P(B) &= P(B \cap A_1) + \dots + P(B \cap A_k) \\ &\text{edo} \\ P(B) &= P(B|A_1)P(A_1) + \dots + P(B|A_k)P(A_k) \end{aligned}$$

Frogapena. Idatz dezagun B gertaera beste modu batera,

$$\begin{aligned} B &= B \cap \Omega \\ &= B \cap \left(\bigcup_{i=1}^{k} A_i \right) \\ &= \bigcup_{i=1}^{k} (B \cap A_i)\,. \end{aligned}$$

$A_i \cap A_j = \emptyset \quad (i \neq j)$ denez, $(B \cap A_i) \cap (B \cap A_j) = \emptyset \quad (i \neq j)$ da, eta, horrela, B gertaera k gertaera bateraezinen bildura gisa adieraztea lortu dugu. Beraz,

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i).$$

Hemendik, PKT aplikatuz, $P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$. $\square$

5.7 adibidea. Komunikabide-kanal bitarrean 0 edo 1 motako seinaleak bidaltzen dira. Zarata dela eta, 0 seinalea bidali baina 1 seinalea jaso daiteke, eta alderantziz. Demagun 0 seinalea jasotzeko probabilitatea benetan 0 bidali denean 0.94 dela, eta 1 seinalea jasotzeko probabilitatea benetan 1 bidali denean 0.91. Gainera, 1 seinalea 0 seinalea baino 11/9 aldiz gehiagotan bidaltzen da. Seinale bat bidali delarik, kalkula dezagun 1 seinalea jasotzeko probabilitatea.

Ebazpena

Defini ditzagun erabiliko ditugun gertaerak.

- $J_0 \equiv$ *«0 seinalea jaso»*
- $J_1 \equiv$ *«1 seinalea jaso»*
- $B_0 \equiv$ *«0 seinalea bidali»*
- $B_1 \equiv$ *«1 seinalea bidali»*

Ematen diguten informazioaren arabera,

$$P(J_0|B_0) = 0.94, \quad P(J_1|B_1) = 0.91.$$

$$\left.\begin{array}{l} P(B_1) = \frac{11}{9} P(B_0) \\ P(B_1) + P(B_0) = 1 \end{array}\right\} \Rightarrow \quad \begin{array}{l} P(B_0) = 0.45 \\ P(B_1) = 0.55 \end{array}$$

$\{B_0, B_1\}$ g.s.o. da eta probabilitate osoaren teorema (POT) aplika dezakegu.

Honela,

$$\begin{aligned} P(J_1) &= P(J_1|B_1)P(B_1) + P(J_1|B_0)P(B_0) \\ &= 0.91 \times 0.55 + 0.06 \times 0.45 \\ &= 0.5275. \end{aligned}$$

Bayes-en teorema. Izan bedi $\{A_1, \dots, A_k\}$ g.s.o $(P(A_i) > 0 \ , i = 1, \dots, k)$. Orduan, $\forall B \in \mathcal{P}(\Omega)$,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B|A_1)P(A_1) + \dots + P(B|A_k)P(A_k)} \ , \quad j = 1 \dots k.$$

Frogapena. Egin ezazu zeuk frogapena. Nahikoa izango duzu lerrotxo bat.

5.8 adibidea. Aurreko adibidearen (5.7) jarraipena.

Kalkula ditzagun:

1. 1 seinalea bidali izateko probabilitatea, 1 seinalea jaso delarik.
2. Errorearen probabilitatea.

Ebazpena

1. Bayesen teorema aplikatuz eta dagoeneko $P(J_1) = 0.5275$ dela kalkulatu dugunez:

$$\begin{aligned} P(B_1|J_1) &= \frac{P(J_1|B_1)P(B_1)}{P(J_1)} \\ &= 0.9488 \end{aligned}$$

2. «Errorea» $= (B_0 \cap J_1) \cup (B_1 \cap J_0)$ kalkulatzeko, biltzen ari garen bi gertaera horiek bateraezinak direla erreparatu behar dugu. Honela,

$$\begin{aligned} P(\text{`Errorea'}) &= P(B_0 \cap J_1) + P(B_1 \cap J_0) \\ &= P(J_1|B_0)P(B_0) + P(J_0|B_1)P(B_1) \\ &= 0.0765 \end{aligned}$$

5.7 Independentzia

Aurreko ataletan ikusi dugu esperimentuan gertatu denaren informazioa edukitzeak gertaeren probabilitateak alda ditzakeela. Hala ere, pentsa dezagun honako kasu honetan kasuan:

> «Txanpon orekatu bat bi aldiz botako dugu». Lehenengo txandan aurpegia atera bada (A_1), zein da bigarren txandan berriro aurpegia (A_2) ateratzeko probabilitatea?

Kasu honetan, argi dago lehenengo aldiz ateratakoak ez duela inongo informaziorik eskaintzen 2.enean zer aterako den «antzemateko». Hau da,

$$\begin{aligned} A_1 &\equiv \textit{Lehen txaponean aurpegia.} \\ A_2 &\equiv \textit{Bigarren txaponean aurpegia.} \end{aligned}$$

Egora horretan, $P(A_1) = 0.5$, $P(A_2) = 0.5$ eta $P(A_2|A_1) = 0.5$. Kasu horretan, azpimarratzekoa da $P(A_2) = P(A_2|A_1)$ dugula. Horrek esan nahi du, lehen txanponarekin gertatutakoak ez duela informaziorik eskaintzen bigarren txanponaren emaitzarentzat.

Horrek guztiak *independentzia estokastiko* kontzeptura garamatza. Intuitiboki, gertaera batek beste gertaera baten inguruko informazioarik eskaintzen ez duenean, gertaera horiek elkarrekiko askeak edo independenteak direla esango dugu.

5.8 definizioa. Izan bitez A eta B gertaerak. A eta B *elkarrekiko askeak* direla esango dugu, baldin eta

$$P(A \cap B) = P(A)P(B) \text{ bada.}$$

5.3 propietatea. *Honako puntu hauek baliokideak dira:*

1. A *eta* B *askeak dira.*
2. $P(A|B) = P(A)$
3. $P(B|A) = P(B)$
4. A *eta* $\bar{B}$ *askeak dira.*
5. $\bar{A}$ *eta* B *askeak dira.*
6. $\bar{A}$ *eta* $\bar{B}$ *askeak dira.*

Frogapena. Baliokidetasun denak frogatzeko, honako kate hau osatuko dugu.

- 1) $\Rightarrow$ 2) frogapena.

$$\begin{aligned} P(A|B) &\overset{def.}{=} \frac{P(A \cap B)}{P(B)} \\ &\overset{1)}{=} \frac{P(A)P(B)}{P(B)} \\ &= P(A)_{\square}. \end{aligned}$$

- 2) $\Rightarrow$ 3)

$$\begin{aligned} P(B|A) &= P(B|A)\frac{P(A|B)}{P(A|B)} \\ &\overset{def.,2)}{=} \frac{P(A\cap B)}{P(A)}\frac{P(A)}{P(A\cap B)/P(B)} \\ &= P(B)_{\square}. \end{aligned}$$

- 3) $\Rightarrow$ 4)

  Batetik, 3) baieztapena erabiliz, $P(\bar{B}|A) = P(\bar{B})$ dugu jarraian ikus daitekeenez.

$$\begin{aligned} P(\bar{B}|A) &= 1 - P(B|A) \\ &\overset{3)}{=} 1 - P(B) \\ &= P(\bar{B}). \end{aligned}$$

  Emaitza hori erabiliz, A eta $\bar{B}$ gertaerak askeak diren iritsiko gara honela:

$$\begin{aligned} P(A\cap\bar{B}) &\overset{def.}{=} P(A)P(\bar{B}|A) \\ &= P(A)P(\bar{B})_{\square}. \end{aligned}$$

- 4) $\Rightarrow$ 5)

  $A = (A\cap B)\cup(A\cap\bar{B})$ eta $B = (B\cap A)\cup(B\cap\bar{A})$ gertaera bateraezinen bildura gisa idatz daitekeenez, $P(A) = P(A\cap B) + P(A\cap\bar{B})$ eta $P(B) = P(A\cap B) + P(\bar{A}\cap B)$ berdintzak ditugu. Bi berdintza horiek bateratuz:

$$\begin{aligned} P(\bar{A}\cap B) &= P(B) - P(A) + P(A\cap\bar{B}) \\ &\overset{4)}{=} P(B) - P(A) + P(A)P(\bar{B}) \\ &= P(B) - P(A) + P(A)(1 - P(B)) \\ &= P(B) - P(A)P(B) \\ &= P(B)(1 - P(A)) \\ &= P(\bar{A})P(B)_{\square}. \end{aligned}$$

- 5) $\Rightarrow$ 6)

  $\bar{A}$ gertaera bateraezinen bildura gisa idatziz, $P(\bar{A}) = P(\bar{A}\cap B) + P(\bar{A}\cap\bar{B})$

dugu. Hortik:

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P(\bar{A}) - P(\bar{A} \cap B) \\ &\overset{5)}{=} P(\bar{A}) - P(\bar{A})P(B) \\ &= P(\bar{A})(1 - P(B)) \\ &= P(\bar{A})P(\bar{B})_{\square}. \end{aligned}$$

- $6) \Rightarrow 1)$

  $\bar{A}$ eta B gertaera bateraezinen bildura gisa egoki idatziz gero, $P(\bar{A}) = P(\bar{A} \cap B) + P(\bar{A} \cap \bar{B})$ eta $P(B) = P(A \cap B) + P(\bar{A} \cap B)$ berdintzak ditugu. Bi berdintza horiek bateratuz:

$$\begin{aligned} P(A \cap B) &= P(B) - P(\bar{A}) + P(\bar{A} \cap \bar{B}) \\ &\overset{6)}{=} P(B) - P(\bar{A}) + P(\bar{A})P(\bar{B}) \\ &= P(B) - P(\bar{A}) + P(\bar{A})(1 - P(B)) \\ &= P(B) - P(\bar{A})P(B) \\ &= P(A)P(B)_{\square}. \end{aligned}$$

Oharra: ez nahastu A eta B askeak izatea eta A eta B bateraezinak izatea.



Pentsa ezazu...

Demagun A eta B bateraezinak direla. Izan daitezke A eta B askeak? Nola?

5.9 adibidea. Herrialde batean, enpresen Interneterako sarbide mota aztertu da. 10 langile edo gehiago dituztenek % 25 osatzen dute, eta horietatik % 60 WIFI bidez sartzen da sarean. 1-9 langile dituztenen artean, aldiz, % 20k besterik ez du erabiltzen WIFIa. Sarerako sarbide mota enpresaren langile kopuruarekiko askea al da?

Ebazpena

$$\begin{aligned} A &\equiv \textit{«WIFI bidezko sarrera»} \\ H &\equiv \textit{«Enpresak 10 langile edo gehiago ditu»} \end{aligned}$$

Ariketak esan duenez, $P(H) = 0.25$, $P(A|H) = 0.60$ eta $P(A|\bar{H}) = 0.20$.

Askeak diren ala ez erantzuteko, $P(H \cap A) \stackrel{?}{=} P(H)P(A)$ betetzen den ala ez ikusi beharko dugu.

Batetik,

$$\begin{aligned} P(H \cap A) &= P(H)P(A|H) \\ &= 0.25 \times 0.60 \\ &= 0.15 \end{aligned}$$

Bestetik,

$$\begin{aligned} P(A) &= P(A|H)P(H) + P(A|\bar{H})P(\bar{H}) \\ &= 0.25 \times 0.60 + 0.75 \times 0.20 \\ &= 0.30 \end{aligned}$$

eta, ondorioz, $P(H)P(A) = 0.25 \times 0.30 = 0.075$.

Beraz, $P(H \cap A) \neq P(H)P(A)$ dugunez, bi gertaera horiek ez dira askeak.

Gainera, $P(A|H) = 0.60 > 0.30 = P(A)$ denez, enpresa handietan WIFIa erabiltzeko probabilitatea handiagoa da oro har dena baino.

5.7.1 Sistemak

Jarraian duzun 5.3 irudian bi sistema irudikatu dira. Batean, (a), osagaiak seriean konektatu dira, eta bestean, (b), paraleloan. Gure helburua da sistemak ongi funtzionatzeko probabilitatearen kalkulua, hau da, sistemaren *fidagarritasuna* zein den kalkulatzea. Horretarako, osagaien funtzionamendua elkarrekiko askea dela onartuko dugu.

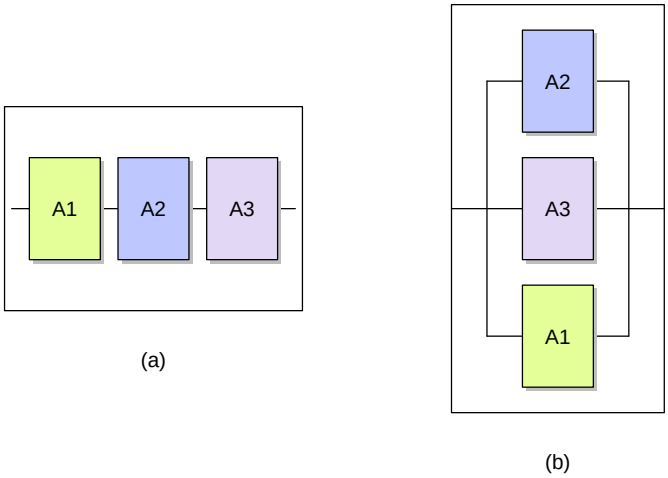
- n **osagai seriean**

  Osagai denak seriean konektaturik badaude, sistema ongi ibiltzeko osagai bakoitzak ongi ibili behar du.

  Izan bedi $A_i \equiv$ «*i. osagaia ongi dabil*», $\quad i = 1, \ldots, n$.

  Orduan,

$$\begin{aligned} P(A_1 \cap \ldots \cap A_n) &= P(A_1) \ldots P(A_n) \\ &= \prod_{i=1}^{n} P(A_i) \end{aligned}$$



5.3 irudia. Seriean (a) eta paraleloan (b) konektaturiko osagaiak.

Adibidez, $P(A_i) = 0.970$ bada, $i = 1, \ldots, n$ eta $n = 5$, sistemaren fidagarritasuna $0.970^5 = 0.859$ra jaitsi da.

n osagai seriean konektaturiko sistema harturik eta $P(A_i) = 0.970$, $i = 1, \ldots, n$ ezarririk, osa ezazu honako taula hau:

n	Fidagarritasuna
2	
5	
10	
50	

Irudika itzazu grafiko batean n eta dagokion fidagarritasuna. Zein ondorio ateratzen duzu?

- n **osagai paraleloan**

  Osagai denak paraleloan konektaturik badaude, sistema ongi ibiltzeko nahikoa da osagai bat ongi ibiltzea.

  Izan bedi $A_i \equiv$ *«i. osagaia ongi dabil»*, $\quad i = 1, \ldots, n$.

Orduan,

$$\begin{aligned} P(A_1 \cup ... \cup A_n) &= 1 - P(\bar{A}_1 \cap ... \cap \bar{A}_n) \\ &= 1 - \prod_{i=1}^{n} P(\bar{A}_i) \\ &= 1 - \prod_{i=1}^{n} (1 - P(A_i)) \end{aligned}$$

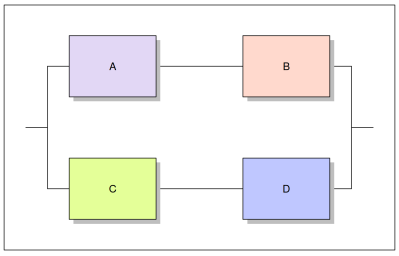
Adibidez, $P(A_i) = 0.970$ bada, $i = 1, ..., n$ eta $n = 5$, sistemaren fidagarritasuna $1 - (1 - 0.970)^5 = 0.999999975$era handitu da.

n osagai paraleloan konektaturiko sistema harturik eta $P(A_i) = 0.970$, $i = 1, ..., n$ ezarririk, osa ezazu honako taula hau:

n	Fidagarritasuna
2	
5	
10	

Irudika itzazu grafiko batean n eta dagokion fidagarritasuna. Zein ondorio ateratzen duzu?

5.10 adibidea. Sistema bat 4 elementuz osatuta dago, azpiko irudian ikus daitekeen moduan.



Osagai bakoitzak ongi funtzionatzeko probabilitatea 0.90 da, eta osagaien funtzionamendua elkarrekiko askea da. Zein da sistema horren fidagarritasuna?

.

Ebazpena

Jar ditzagun $A \equiv$ *«A osagaia ongi dabil»* eta berdin B, C eta D gertaeretarako. Orduan, sistemaren fidagarritasuna honako hau da:

$$p \quad = \quad P((A \cap B) \cup (C \cap D)).$$

Dei diezaiegun $S_1 = A \cap B$ eta $S_2 = C \cap D$. Osagaien independentzian oinarrituz, $P(S_1) = P(S_2) = 0.9^2$ eta $P(S_1 \cap S_2) = 0.9^4$.

Fidagarritasunaren kalkulura itzuliz:

$$\begin{aligned} p &= P((A \cap B) \cup (C \cap D)) \\ &= P(S_1) + P(S_2) - P(S_1 \cap S_2) \\ &= 2 \times 0.9^2 - 0.9^4 \\ &= 0.9639. \end{aligned}$$

6

Probabilitate-kalkulua: zorizko aldagaiak eta probabilitate-banaketak

6.1 Zorizko aldagaiaren ideia orokorra

Zorizko aldagaiaren ideia ulertzeko, adibide batekin hasiko gara.

6.1 adibidea. Honako joko hau proposatu zaio jokalariari: dadoa botako dugu, eta jokalariak 10 euro irabaziko ditu baldin eta 6a ateratzen bada. Ez bada, ordea, 6a ateratzen, jokalariak 2 euro galduko ditu.

Saiatu gaitezen jokalariari gerta dakiokeena formalki azaltzen. Jokalariari gerta dakiokeena X-z adieraziko dugu. Jokalariari bi gauza gerta dakizkioke: $X = 10$ edo $X = -2$, eta hori zuzenean erlazionatuta dago dadoaren emaitzarekin. Horregatik, ezingo dugu aurresan zer gertatuko den; hau da, zorizkoa da, baina hala ere zehaztu dezakegu zer gerta daitekeen eta baita zein probabilitaterekin gerta daitekeen ere.

Ω izanik dadoa botatzearen lagin-espazioa eta $M = \{-2, 10\}$ jokalariari gerta dakiokeen emaitzen multzoa, bi horien arteko erlazioa honela adieraz dezake-

gu:

$$\begin{aligned} X \quad : \quad & \Omega \to M \\ & 1 \hookrightarrow -2 \\ & 2 \hookrightarrow -2 \\ & 3 \hookrightarrow -2 \\ & 4 \hookrightarrow -2 \\ & 5 \hookrightarrow -2 \\ & 6 \hookrightarrow \ 10 \end{aligned}$$

Hau da,

$$X: \begin{array}{ll} \Omega \to & M \\ \omega \hookrightarrow & X(\omega) \end{array} \qquad \text{eta} \qquad X(\omega) = \begin{cases} 10, & \omega = 6, \\ -2, & \omega \neq 6. \end{cases}$$

Funtzio horren aurreirudiari begiratzen badiogu, $X^{-1}(-2) = \{1, 2, 3, 4, 5\}$ eta $X^{-1}(10) = \{6\}$ multzoak ditugu. Multzo horiek $\mathcal{P}(\Omega)$-ren elementuak dira, eta $(\Omega, \mathcal{P}(\Omega), P)$ probabilitate-espazioa denez, naturalki ezar ditzakegu:

$$\begin{aligned} P(X = -2) = P(X^{-1}(-2)) = P(\{1, 2, 3, 4, 5\}) = 5/6 \text{ eta} \\ P(X = 10) = P(X^{-1}(10)) = P(\{6\}) = 1/6. \end{aligned}$$

Jokalariari gerta dakiokeena aztertzeko, esperimentuaren emaitzetan baino gehiago, horri loturiko X funtzioan dago gure interesa. Zorizko esperimentuaren gainean definiturik dagoen probabilitate-funtzioak zuzenean ezartzen du X-ren balioen gaineko probabilitatea. X-ri *zorizko aldagaia* esango diogu.

6.1 definizioa. Izan bedi E zorizko esperimentua eta Ω honi dagokion lagin-espazio finitua. Defini dezagun honela X:

$$X: \begin{array}{ll} \Omega \to & M \subseteq \mathbb{R} \\ \omega \hookrightarrow & X(\omega) \end{array}$$

Itxura horretako funtzioei *zorizko aldagaia* esaten zaie. M zorizko aldagaiaren *behaketa espazioa* da.

Oharra: askotan, z.a. soilik idatziko dugu zorizko aldagaia adierazteko.

Baldin eta P probabilitate-funtzioa badugu zorizko esperimentuari loturik, hau da, $(\Omega, \mathcal{P}(\Omega), P)$ probabilitate-espazioa bada, $P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega | X(\omega) = x\})$ ulertuko dugu.

Zorizko aldagaiak hartu ditzakeen balioen arabera, *diskretuak* eta *jarraituak* bereizten dira:

- **Diskretuak:** aldagaiak hartu ditzakeen balioek (probabilitate positiboarekin) kopuru finitua edo zenbagarria osatzen dute.
- **Jarraituak:** aldagaiaren behaketa espazioak tarte erreal bat barneratzen duenean (probabilitate positiboarekin).

6.2 Probabilitate-legea eta dentsitate-funtzioa

6.2 definizioa. Izan bedi X z.a. diskretua. $f(x_i) = P(X = x_i)$ balioek ($i = 1, 2, \dots, I, \dots$) X-ren *probabilitate-legea* osatzen dute, baldin eta:

i. $f(x_i) \geq 0\,, \quad \forall i.$
ii. $\sum_i f(x_i) = 1.$

Definizioan $M = \{x_1, x_2, \dots, x_I, \dots\}$ infinitua jarri badugu ere, erreparatu kasu finitua ere barneratu dezakeela definizioak, $P(X = x_i) = 0$ ezarriz $i = I{+}1, I{+}2, \dots$ indizeetarako. Terminologiari dagokionez, sarri, probabilitate-legea *probabilitate-masez* osatuta dagoela esaten da.

6.2 adibidea. Pieza bakoitza ona (1) ala txarra (0) izan daiteke, eta piezak 3ko sortetan iristen zaizkigu. Zoriz sorta bat hartu, eta sortan dagoen pieza onen kopurua interesatzen zaigu. Zein dira egoera honetan sortzen zaigun z.a. eta bere probabilitate-legea?

Ebazpena

Dei deizaiogun $X \equiv$«Sortan dagoen pieza on kopurua»ri. Nolako balioak hartu ditzakeen ikusteko, zorizko espermentuaren zuhaitz-diagrama eraikiko dugu:

1. pieza	2. pieza	3. pieza	Lagin-esp. Ω	X
1	1	1	(1, 1, 1)	3
1	1	0	(1, 1, 0)	2
1	0	1	(1, 0, 1)	2
⋮	⋮	⋮	⋮	⋮
0	0	1	(0, 0, 1)	1
0	0	0	(0, 0, 0)	0

Zuhaitz-diagraman ikusten dugu 8 oinarrizko gertaera daudela ($\#\Omega = 8$), gertaera horiek ekiprobableak direla, eta X-ren 4 balio desberdinekin lotzen direla: 0, 1, 2 edo 3 balioak; hain zuzen ere. 0 eta 3 balioei dagokienez, bakoitza gertaera bakarrarekin lotzen da, eta 1 eta 2 balioei dagokienez, bakoitza 3 gertaerekin.

Eta, beraz, hau da X-ren probabilitate-legea edo probabilitate-masa:

$$f(x) = \begin{cases} 1/8\,, & x = 0 \\ 3/8\,, & x = 1 \\ 3/8\,, & x = 2 \\ 1/8\,, & x = 3 \end{cases}$$

Probabilitate-legeak erakusten digu, batetik, X-rentzako balio posibleak 0, 1, 2 eta 3 direla; bestetik, zein diren balio horiei dagozkien probabilitateak.

6.3 definizioa. Izan bedi X z.a. jarraitua. Honako baldintza hauek betetzen dituen f funtzio errealari X-ren *dentsitate-funtzioa* esaten zaio.

i. $f(x) \geq 0\,, \quad \forall x.$
ii. $\int_{-\infty}^{\infty} f(x)dx = 1.$

Dentsitate-funtzioak X-ren probabilitate-banaketa neurtzen du; izan ere,

$$P(a \leq X \leq b) = \int_a^b f(x)dx\,, \quad a < b.$$

6.3 adibidea. Demagun $X \equiv$ *«Sistemak erantzuna eman arteko denbora segundotan»* dugula eta hau dela bere dentsitate-funtzioa:

$$f(x) = \begin{cases} \frac{1}{5}e^{-\frac{1}{5}x}\,, & x \geq 0 \\ 0, & \text{bestela} \end{cases}$$

Kalkula dezagun sistemari eskaera egin eta 5 segundo baino gehiago itxaroteko probabilitatea.

Ebazpena

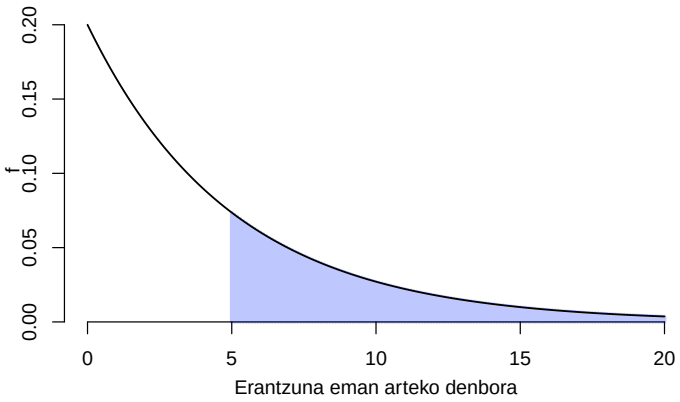
Formalki, $P(X \geq 5)$ kalkulatu behar dugu:

$$\begin{aligned} P(X \geq 5) &= \int_5^{\infty} \frac{1}{5}e^{-\frac{1}{5}x}dx \\ &= -e^{-\frac{1}{5}x}\Big]_5^{\infty} \\ &= 0 + e^{-1} \\ &= 0.368 \end{aligned}$$

Beraz, eredu horren arabera, sistemari eskaera egin eta 5 segundo edo gehiago

itxaroteko probabilitatea 0.368 da.

Hurrengo irudian, egindako kalkulua grafikoki adierazten da. Zehazki, irudiak X-ren dentsitate-funtzioa erakusten du, eta ilunduta azaltzen den azalerak $P(X \geq 5)$ probabilitatea irudikatzen du.



Oharra: zorizko aldagaia *jarraitua* denean,

$$\begin{aligned} P(a \leq X \leq b) &= P(a \leq X < b) \\ &= P(a < X \leq b) \\ &= P(a < X < b). \end{aligned}$$

6.3 Banaketa-funtzioa

6.4 definizioa. Izan bedi X z.a. X-ren *banaketa-funtzioa* honela definitzen da:

$$\begin{aligned} F \quad : \quad & \mathbb{R} \to [0,1] \\ & x \hookrightarrow F(x) = P(X \leq x) \end{aligned}$$

Probabilitate-legeak edo dentsitate-funtzioak ematen duen informazioa eta banaketa-funtzioak ematen duena baliokideak dira. Hau da, f ezagututa F kalkula dezakegu, eta baita alderantziz ere. Askotan, X z.a. nolakoa den ezagutzeko probabilitate-legearen edo dentsitate-funtzioaren ordez *banaketa-funtzioa* erabiltzen da. Zenbait kasutan, errazagoa izan daiteke zorizko aldagaiari dagokion F kalkulatzea f kalkulatzea baino.



Pentsa ezazu...

Banaketa-funtzioaren definizioan esaten da nola kalkulatu F, f probabilitate-legea edo dentsitate-funtzioa ezagututa. Nola egingo zenuke, ordea, alderantziz, hau da, nola kalkultatuko zenuke f banaketa-funtzioa ezagututa?

6.4 adibidea. X zorizko aldagai diskretuaren banaketa-funtzioa

Kalkula dezagun honako probabilitate-lege hau duen z.a.-ren banaketa funtzioa, (6.2) adibidetik datorrena.

$$f(x) = \begin{cases} 1/8\,, & x = 0 \\ 3/8\,, & x = 1 \\ 3/8\,, & x = 2 \\ 1/8\,, & x = 3 \end{cases}$$

Ebazpena

Zuzen erreal osoan definiturik dagoen F banaketa-funtzioa eraiki behar dugu. Finka dezagun $x \in \mathbb{R}$ edozein balio.

- Baldin eta $x < 0$ bada, $P(X \leq x) = 0$ izango dugu.
- Finkaturiko x balioa $0 \leq x < 1$ bada, $P(X \leq x) = P(X = 0) = 1/8$.
- Baldin eta $1 \leq x < 2$ bada, $P(X \leq x) = P(X = 0) + P(X = 1) = 1/2$.
- Baldin eta $2 \leq x < 3$ bada, $P(X \leq x) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8$.
- Baldin eta $x \geq 3$ bada, $P(X \leq x) = 1$.

Beraz, edozein delarik $x \in \mathbb{R}$, bere $F(x)$ balioa kalkulatu dugu. Laburbilduz:

$$F(x) = \begin{cases} 0\,, & x < 0 \\ 1/8\,, & 0 \leq x < 1 \\ 1/2\,, & 1 \leq x < 2 \\ 7/8\,, & 2 \leq x < 3 \\ 1\,, & x \geq 3 \end{cases}$$

6.5 adibidea. X **zorizko aldagai jarraituaren banaketa-funtzioa**

Lehentxeago (6.3) adibidean ikusi dugunez, $X \equiv$ *«Sistemak erantzuna eman arteko denbora segundotan»* harturik, bere dentsitate-funtzioa honako hau zen:

$$f(x) = \begin{cases} \frac{1}{5}e^{-\frac{1}{5}x}, & x \geq 0 \\ 0, & \text{bestela} \end{cases}$$

Kalkula dezagun zein den bere banaketa-funtzioa.

Ebazpena

Kasu honetan, dentsitate-funtzioak zuzen errealeko bi eremu bereizten ditu:

- $x < 0$ denean, $F(x) = P(X \leq x) = 0$.
- $x \geq 0$ denean,

$$\begin{aligned} F(x) = P(X \leq x) &= \int_0^x \frac{1}{5}\exp\{-\frac{1}{5}t\}dt \\ &= -\exp\{-\frac{1}{5}t\}\Big]_0^x = -\exp\{-\frac{1}{5}x\} + 1 \end{aligned}$$

Hau da,

$$F(x) = \begin{cases} 0 \quad , & x < 0 \\ 1 - e^{-\frac{1}{5}x}, & x \geq 0 \end{cases}$$

6.4 Itxaropena eta bariantza

Orain, ezagutu behar ditugun kontzeptuak lantzeko adibide batekin hasiko gara:

6.6 adibidea. Kasinora joan nahi dugu ruletan jokatzera. Ruletara jokatzeko, 1etik 36rako zenbaki bat aukeratu behar dugu, eta diru bat jarri; demagun 10 euro. Ruleta bota eta guk aukeraturiko zenbakia ateratzen bada, guk jarritako kopurua 36 aldiz itzuliko digute. Aldiz, ez bada guk aukeratutako zenbakia ateratzen, jarritako dirua galduko dugu. Jokoaren mekanismoa hori izanik, badakigu: batetik, gehienetan galdu egingo dugula eta oso gutxitan irabazi; baina, bestetik, baita ere badakigu galtzen dugunean gutxi galduko dugula eta, aldiz, irabazten dugunean asko irabaziko dugula. Beraz, jakin nahi duguna honako hau da: *batez beste, partidako zenbat diru irabaztea espero dugu?*

Ebazpena

Azter dezagun nolakoa den $X \equiv$ *«Irabazia»* gure zorizko aldagaia. Demagun $i \in \{1, ..., 36\}$ zenbakira egin dugula apustua. Orduan, lagin-espazioko elementuak eta zorizko aldagaiaren elementuak honela erlazionatuko lirateke:

$$\begin{array}{rccccccc} \Omega = \{ & 0, & 1, & \dots, & i, & \dots, & 36\} \\ & \downarrow & \downarrow & & \downarrow & & \downarrow \\ X: & -10 & -10 & & 36 \times 10 - 10 & & -10 \end{array}$$

Hortik, X-ren probabilitate-legea honako hau da:

$$f(x) = \begin{cases} 36/37\,, & x = -10 \\ 1/37\,, & x = 350 \end{cases}$$

Gure helburua *batez beste, partidako zenbat diru irabaztea espero dugu?* galdera erantzutea izanik, suposa dezagun kasinora joatea erabaki dugula zer gerta dakigukeen ikusteko. Hau da, demagun 50 aldiz **jokatu ondoren**, behin irabazi ditugula 350 euro eta gainontzeko 49 apostutan galdu dugula. Egoera horretan, erraza da kalkulatzea batez beste zenbat irabazi dugun partidako:

$$\begin{aligned} \frac{350 + (-10) \times 49}{50} &= -2.8 \\ 350 \times \frac{1}{50} + (-10) \times \frac{49}{50} &= -2.8 \end{aligned} \tag{6.1}$$

Kasu horretan, ez diogu emaitzari begiratuko, baizik eta emaitza lortzeko idatzi dugun (6.1) adierazpenari. Adierazpen horren ezker aldea begiratzen badugu, esaten ari dena zera da: 350 balioa ikusi dugu 1/50 maiztasunarekin, eta -10 balioa 49/50 maiztasunarekin. X aldagaiak hartu ditzakeen balioen batezbesteko haztatua dugu.

Hori ikusirik, kasinoan **jokatu aurretik** honako kalkulu hau egin dezakegu. Badakigu, $X = 350$ balioa $P(X = 350) = 1/37$ probabilitateaz emango dela, eta $X = -10$ balioa, berriz, $P(X = -10) = 36/37$ probabilitateaz. Hortik, ruletara jokatzerakoan, batez beste,

$$350 \times \frac{1}{37} + (-10) \times \frac{36}{37}$$

kopurua irabaztea espero dezakegu. Hau da, $350 \times f(350) + (-10) \times f(-10) = -0.27 < 0$ irabaztea espero dugu partidako.

Emaitza gisa, errepara dezagun negatiboa dela, hots, batez beste partidako galtzea espero behar dudala (kasinoak, ordea, irabaztea!). Baina, batez ere,

errepara dezagun kasinoan jokatu aurretik kalkulatu dugula eta zein arrazonamendurekin kalkulatu dugun; izan ere, zuzenean *itxaropen matematikoa* kontzeptura garamatza.

6.5 definizioa. Izan bedi X zorizko aldagaia. X-ren *itxaropen matematikoa* honela definitzen da:

i. X z.a. diskretua bada:
$$E(X) = \sum_x xf(x)$$

ii. X z.a. jarraitua bada:
$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$E(X) < \infty$ bada, $E(X) \in \mathbb{R}$ zenbaki bat da, X-ren balioen «erdigunean» kokatuko dena.

6.7 adibidea. Hartu dezagun berriki ikusi dugun adibidea: $X \equiv$ *«Sistemak erantzuna eman arteko denbora segundotan»* zorizko aldagaia eta bere dentsitate-funtzioa:
$$f(x) = \begin{cases} \frac{1}{5}e^{-\frac{1}{5}x}, & x \geq 0 \\ 0, & \text{bestela} \end{cases}$$

Orain kalkulatu nahi duguna da, batez beste, zenbat itxaron behar dugun sistemak erantzuna eman arte, eredu horren arabera. Hau da, $E(X) = ?$

Ebazpena

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} x\frac{1}{5}e^{-\frac{1}{5}x}dx = \begin{bmatrix} u = x\,, & du = dx \\ dv = 1/5e^{-1/5\,x}\,, & v = -e^{-1/5\,x} \end{bmatrix}$$
$$= -xe^{-\frac{1}{5}x}\Big]_0^{\infty} + \int_0^{\infty} e^{-\frac{1}{5}x}dx =$$
$$= -5e^{-\frac{1}{5}x}\Big]_0^{\infty} = 5\,.$$

Eredu horren arabera, sistemari eskaera egin eta, batez beste, 5 segundo itxaron behar izatea espero da.

6.1 propietatea. *Izan bedi X z.a. eta EX bere itxaropena.*

1. $\forall a, b \in \mathbb{R}, E(aX + b) = aEX + b$

2. *Izan bedi* $c \in \mathbb{R}$ *konstantea,*

$$E(c) = c$$

3. *Izan bedi* Y *z.a.*

$$E(X+Y) = EX + EY$$

4. *Izan bitez* X *eta* Y *z. a. askeak* [1]

$$E(X \cdot Y) = EX \cdot EY$$

5. *Izan bedi* h *funtzio erreala eta hartu dezagun* $Y = h(X)$ *zorizko aldagai berria.*

   i. $E(Y) = E(h(X)) = \sum_x h(x)f(x)$*,* X *diskretua bada.*
   ii. $E(Y) = E(h(X)) = \int h(x)f(x)dx$*,* X *jarraitua bada.*

6.6 definizioa. Izan bedi X zorizko aldagaia. X-ren *bariantza* honela definitzen da:

i. X z.a. diskretua bada:

$$\begin{aligned} VAR(X) &= \sum_x (x - EX)^2 f(x) \\ &= E\left(X - EX\right)^2 \end{aligned}$$

ii. X z.a. jarraitua bada:

$$\begin{aligned} VAR(X) &= \int_{-\infty}^{\infty} (x - EX)^2 f(x)dx \\ &= E\left(X - EX\right)^2 \end{aligned}$$

Bariantzak X zorizko aldagaiaren sakabanapena neurtzen du.

6.2 propietatea. *Izan bedi* X *z.a.* $EX = \mu$ *delarik. Orduan,*

$$\begin{aligned} VAR(X) &= EX^2 - \left(EX\right)^2 \\ &= EX^2 - \mu^2 \end{aligned}$$

Frogapena. Bariantzaren definizioko espresioa garatuko dugu:

$$\begin{aligned} VAR(X) &= E\left(X - \mu\right)^2 = E\left(X^2 - 2\mu X + X^2\right) \\ &= EX^2 + E(-2\mu X) + E\mu^2 \\ &= EX^2 - 2\mu EX + \mu^2 \\ &= EX^2 - \mu^2. \end{aligned}$$

[1] X eta Y askeak dira, $\forall x, y$ izanik, $P(X = x, Y = y) = P(X = x)P(Y = y)$ bada.

6.3 propietatea. *Izan bedi X z.a.*

1. $VAR(X) \geq 0$
2. $VAR(X) = 0 \quad \Leftrightarrow \quad X = ktea.$
3. $\forall a, b \in \mathbb{R} \quad VAR(aX + b) = a^2 VAR(X)$
4. *Izan bitez X eta Y z. a. askeak*

$$VAR(X + Y) = VAR(X) + VAR(Y)$$

6.7 definizioa. Bariantzaren erro karratuari *desbideratze estandarra* esaten zaio.

$$SD(X) = \sqrt{VAR(X)}$$

6.8 adibidea. $X \equiv$ *«Sistemak erantzuna eman arteko denbora segundotan»* zorizko aldagaia berriro hartuz, kalkula ditzagun zein diren bere bariantza eta desbideratze estandarra.

Ebazpena

Gogora dezagun $EX = 5$ kalkulatu genuela. Bariantza kalkulatzeko nahikoa dugu EX^2 kalkulatzea.

Egiazta ezazu $EX^2 = 50$ dela. Orduan, $VAR(X) = EX^2 - (EX)^2 = 50 - 5^2 = 25$ eta $SD = \sqrt{25} = 5$ dira.

6.5 Hainbat banaketa ezagun

Jarraian, hainbat probabilitate-banaketa ezagunen azalpenak eta definizioak dituzu. Lehenengo, aldagai diskretuei dagozkienak topatuko dituzu, eta, ondoren, aldagai jarraituei dagozkienak.

6.5.1 Bernoulli-ren banaketa

Demagun zorizko esperimentuak bi emaitza posible bakarrik dituela: *arrakasta* ala *porrota*. Eraiki dezagun X zorizko aldagaia, 1 ala 0 balioak esleituz esperimentuaren emaitzaren arabera. Hau da, $X = 1$ izango da esperimentuan *arrakasta* lortu denean eta $X = 0$ esperimentuan *porrota* lortu denean.

Demagun esperimentuan *arrakasta* lortzeko probabilitatea p dela. Orduan, X-ren probabilitate-legea honako hau da:

Probabilitate-legea:

$$\begin{aligned} P(X=x) &= \begin{cases} p\ , & x=1 \\ 1-p\,, & x=0 \end{cases} \\ &= p^x(1-p)^{1-x}\,, \quad x=0,\,1. \end{aligned}$$

$X \sim Bernoulli(p)$ eran adieraziko dugu.

Ezaugarriak:

$$E\,X = p \quad \text{eta} \quad VAR(X) = p(1-p)$$



Pentsa ezazu...

Zein zorizko aldagairekin eta nola adieraziko zenuke *partxisean* dadoa botatzean fitxa «etxetik» ateratzea ala ez ateratzea?

6.5.2 Banaketa binomiala

Banaketa binomiala honako baldintzadun zorizko esperimentu honi lotzen zaio:

- Esperimentuan n saiakera egiten dira, eta n, saiakera kopurua, aurretik finkaturik dago.
- Saiakera guztiak berdinak dira, eta bakoitzean bi emaitza posible bakarrik daude:
  - *arrakasta* (1),
  - *porrota* (0).
- Saiakera guztiak elkarrekiko askeak dira.
- Saiakera bakoitzean *arrakasta* lortzeko probabilitatea konstantea da (p deituko diogu).

Izan bedi $X \equiv$ *«n saiakera horietan lortutako arrakasta kopurua»* zorizko aldagaia. Orduan, X-k banaketa binomiala jarraitzen duela esango dugu, eta honela adieraziko:

$$X \sim Bin(n,p)$$

Probabilitate-legea:

$$P(X=x) = \binom{n}{x} p^x(1-p)^{n-x}\,, \quad x=0,1,\ldots,n\,, \text{ delarik.}$$

Ezaugarriak:

$$E\ X = np \quad \text{eta} \quad VAR(X) = np(1-p)$$



Pentsa ezazu...

Kamera digitalen fabrikatzaile batek dio garantia-garaian kameren % 10ak soilik behar duela erreparazioren bat. Denda batek honelako 20 kamera erosi ditu. Zein da gehienez kamera batek garantia-garaian erreparazioren bat behar izateko probabilitatea?

Taulen erabilera

Kalkuluak errazteko **estatistikako softwarea** erabil dezakezu, edo, dagoeneko probabilitateak tabulatuta dauden taulak erabil ditzakezu (liburuxkako A eranskinean jasota dituzu). Demagun $X \sim Bin(9, 0.2)$ dela. Kalkula itzazu honako probabilitate hauek:

1. $P(X \leq 4)$
2. $P(X \geq 3)$
3. $P(X = 7)$

```
# Probabilitate-legea: dbinom()
# Banaketa-funtzioa: pbinom()

# P(X <= 4)
pbinom(4, size=9, prob=0.2)

# P(X >= 3)
1 - pbinom(2, size=9, prob=0.2)

# P(X = 7)
dbinom(7, size=9, prob=0.2)
# edo
pbinom(7, size=9, prob=0.2) - pbinom(6, size=9, prob=0.2)
```

6.5.3 Banaketa hipergeometrikoa

Banaketa hipergeometrikoa honako baldintzadun zorizko esperimentu honi lotzen zaio:

- Populazioa N tamainakoa da ($N < \infty$).

- Populazioko ale bakoitzak *arrakasta* ala *porrot* egokitua izango du. Gainera, badakigu populazioan M alek dutela arrakasta egokitua.
- n tamainako lagina aukeratzen da, zoriz; hau da, populazioko ale guztiek aukeratuak izateko probabilitate berdina dute.

Izan bedi $X \equiv$ *«Laginean dagoen arrakasta kopurua»* zorizko aldagaia. Orduan, X-k banaketa hipergeometrikoa jarraitzen duela esango dugu, eta honela adieraziko:

$$X \sim H(N, M, n)$$

Probabilitate-legea:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = a, a+1, \ldots, b-1, b,$$

$a = max\{0,\ n-(N-M)\}$ eta $b = min\{n, M\}$ direlarik.

Ezaugarriak:

$$E\ X = n\frac{M}{N} \quad \text{eta} \quad VAR(X) = n\frac{N-n}{N-1}\frac{M}{N}\left(1-\frac{M}{N}\right)$$

Erregela:

Izan bedi $X \sim H(N, M, n)$. Populazioaren tamaina laginaren tamainarekiko oso handia denean, hau da, $0.05N > n$ denean, $X \approx Bin(n, \frac{M}{N})$ onar daiteke.



Pentsa ezazu...

Izurrite bat dela eta, osasun-erakunde bateko arduradunen populazioaren % 5a jota dagoela estimatzen dute. Azterketa sakon bat egitearren, zoriz aukeratu dute 200 pertsonaz osatutako lagin bat, 200 000 pertsonaz osatutako populaziotik. Zein da populazioan 10 pertsona gaixo topatzeko probabilitatea?

6.5.4 Banaketa binomial negatiboa

Banaketa binomial negatiboa honako baldintzadun zorizko esperimentu honi lotzen zaio:

- Esperimentua bata bestearen ondorengo saiakeraz osatuta dago. Saiakera horiek elkarrengandik askeak dira.

- Saiakera bakoitzak bi emaitza posible besterik ez ditu: *arrakasta* ala *porrota*.
- Saiakera batean arrakasta lortzeko probabilitatea (p) konstantea da saiakeraz saiakera.
- Esperimentuak guztira r arrakasta lortu arte jarraitzen du.

Izan bedi $X \equiv$ *«r. arrakasta lortu arte izandako porrot kopurua»* zorizko aldagaia. Orduan, X-k binomial negatiboa jarraitzen duela esango dugu, eta honela adieraziko:

$$X \sim BinNeg(r, p)$$

Probabilitate-legea:

$$P(X = x) = \binom{x + r - 1}{x} (1 - p)^x p^r \,, \quad x = 0, 1, 2, ...$$

Ezaugarriak:

$$E\ X = \frac{r(1-p)}{p} \quad \text{eta} \quad VAR(X) = \frac{r(1-p)}{p^2}$$

Oharra: $BinNeg(1, p)$ banaketari, askotan, banaketa *geometrikoa* esaten zaio ($BinNeg(1, p) \equiv Geom(p)$).



Pentsa ezazu...

Sistema elektroniko batean, fabrikatzaileak fusible elektrikoak erabiltzen ditu. Fusibleak multzo handitan erosten dira, eta lehenengo akastuna topatu arte frogatzen dira. Demagun multzoko % 10a akastuna dela. Kalkula ezazu lehen akastuna frogatutako lehen 5 fusibleen artean egoteko probabilitatea.

6.5.5 Poissonen banaketa

Banaketa hau ez zaio zehazki esperimentu konkretu bati lotzen. Baina esan daiteke maiz egokia suertatzen dela neurri jakineko denbora-tarteetan gertaera bat izatea «zaila» denean gertaera kopurua zenbatzen duen zorizko aldagaiak Poissonen banaketari jarraitzea.

X zorizko aldagaiak λ parametrodun Poissonen banaketari jarraitzen diola esango dugu, baldin eta bere probabilitate-legea honako hau bada:

Probabilitate-legea:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, ...$$

$X \sim \mathcal{P}(\lambda)$ moduan adieraziko dugu.

Ezaugarriak:

$$E\ X = \lambda \quad \text{eta} \quad VAR(X) = \lambda$$



Pentsa ezazu...

Ordenagailu-sare batean sistema eroritako aldi kopuruak 1 parametrodun Poissonen banaketari jarraitzen dio. Zein da sistema behin baino gehiagotan erortzeko probabilitatea?

6.5.6 Banaketa uniformea

X zorizko aldagaiak a eta b parametrodun banaketa uniformeari jarraitzen diola esango dugu, baldin eta bere **dentsitate-funtzioa** honako hau bada:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{bestela} \end{cases}$$

$X \sim \mathcal{U}(a, b)$ moduan adieraziko dugu.

Ezaugarriak:

$$E\ X = \frac{a+b}{2} \quad \text{eta} \quad VAR(X) = \frac{(b-a)^2}{12}.$$



Pentsa ezazu...

Demagun egunero autobusez joaten garela fakultatera eta autobusak 5 minuturo izaten direla. Beti ordu berean zehazki ez gara iristen autobusgeltokira; $X \equiv$ *«Itxaron beharreko denbora (min)»* bada, eman ezazu X-rentzako dentsitate-funtzio egoki bat.

6.5.7 Banaketa normala

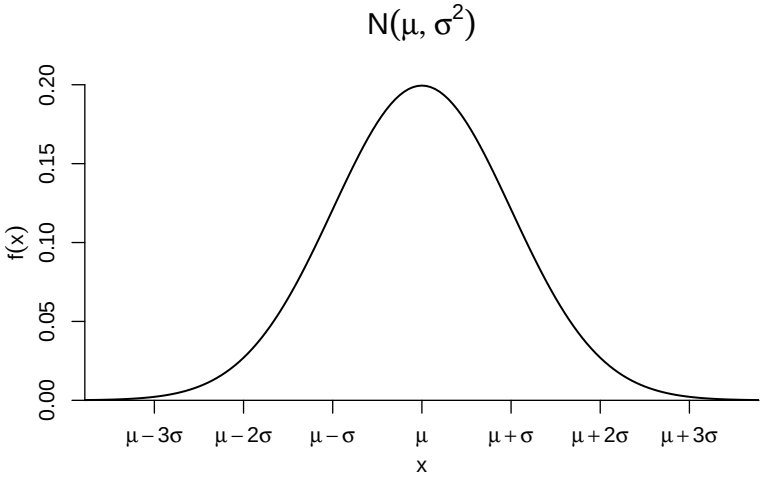
X zorizko aldagaiak μ eta σ^2 parametrodun banaketa normalari jarraitzen diola esango dugu, baldin eta bere **dentsitate-funtzioa** honako hau bada (ikusi 6.1 irudia):

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

$X \sim N(\mu, \sigma^2)$ moduan adieraziko dugu.

Ezaugarriak:

$$E\,X = \mu \quad \text{eta} \quad VAR(X) = \sigma^2.$$



6.1 irudia. Banaketa normala.

6.4 propietatea. *Aldagaia estandarizatzea*

$$X \sim N(\mu, \sigma^2) \quad \Leftrightarrow \quad Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

X aldagaitik Z aldagaia lortzen dugunean, X aldagaia estandarizatu *dugula esaten dugu. $N(0, 1)$ banaketari ere sarri banaketa normal* estandarizatua *esaten zaio.*

Taulen erabilera

Banaketa normalaren dentsitate-funtzioa intuitiboki ulertzea erraza bada ere, dagozkion probabilitate-kalkuluak egitea ez da berehalakoa. Horregatik, banaketa

normalari dagozkion probabilitateak kalkulatzeko **software estastistikoaz** baliatuko gara, edo, bestela, dagoeneko probabilitate horiek kalkulatuta dituzten taulez.

Demagun $Z \sim N(0,1)$ dela. Kalkula itzazu honako probabilitate hauek:

1. $P(Z \leq 2)$
2. $P(Z \geq 2.18)$
3. $P(Z \leq -2)$

```
# Dentsitate-funtzioa: dnorm()
# Banaketa-funtzioa: pnorm()


# P(Z <= 2)
pnorm(2, mean=0, sd=1)

# P(Z >= 2.18)
1 - pnorm(2.18, mean=0, sd=1)

# P(Z <= -2)
pnorm(-2, mean=0, sd=1)
# edo
1 - pnorm(2, mean=0, sd=1)
```

Normalaren taulak erabiltzeko, kontuan izan behar dugu $N(0,1)$ normalaren taula soilik dugula. Horregatik, edozein dela ere zure zorizko aldagaia, taula erabili aurretik estandarizatu egin beharko duzu. Software estatistikoa erabiltzen ari bazara, estandarizazio hori ez da beharrezkoa.

Demagun $X \sim N(10, 2^2)$ dela. Kalkula itzazu honako probabilitate hauek:

1. $P(X \leq 15)$
2. $P(6 \leq X \leq 14)$
3. $P(X \geq 17)$

```
# Dentsitate-funtzioa: dnorm()
# Banaketa-funtzioa: pnorm()


# P(X <= 15)
pnorm(15, mean=10, sd=2)

# P(6 <= X <= 14)
pnorm(14, mean=10, sd=2) - pnorm(6, mean=10, sd=2)

# P(Z >= 17)
1 - pnorm(17, mean=10, sd=2)
```

6.9 adibidea. Demagun hegazkin pilotua 18 000 oinetako altuera mantentzen saiatzen dela, hegazkinaren altuera $\mu = 18\,000$ eta $\sigma^2 = 50^2$ parametrodun zorizko aldagaia dela, eta bere banaketa *normala* dela. Zein da hegazkinaren altuera mantendu beharreko 18 000 oinetatik 100 oinetako diferentzia batera egoteko probabilitatea?

Ebazpena

Gure interesa $X \equiv$ *«Hegazkinaren altuera»* aldagaian dago eta $X \sim N(18\,000, 50^2)$ dela esaten du enuntziatuak. Zehazki, $P(|X - 18\,000| \leq 100)$ kalkulatzeko eskatu du.

$$\begin{aligned} P(|X - 18\,000| \leq 100) &= P(-100 \leq X - 18\,000 \leq 100) \\ &\left[\begin{array}{l} X \text{ estandarizatu behar dugu.} \\ \text{Izan bedi: } Z = \dfrac{X - 18\,000}{50} \\ X \sim N(18\,000, 50^2) \text{ denez, } \Rightarrow Z \sim N(0,1) \end{array}\right] \\ &= P(-2 \leq Z \leq 2) \\ &= P(Z \leq 2) - P(Z \leq -2) \\ &= 0.9545 \end{aligned}$$

6.5.8 Banaketa esponentziala

$X \sim Exp(\lambda)$

Dentsitate-funtzioa: (ikusi 6.2 irudia)

$$f(x) = \begin{cases} \lambda\, e^{-\lambda x}, & x \geq 0 \\ 0 \quad , & \text{bestela} \end{cases}$$

Ezaugarriak:

$$E\,X = \frac{1}{\lambda} \quad \text{eta} \quad VAR(X) = \frac{1}{\lambda^2}.$$

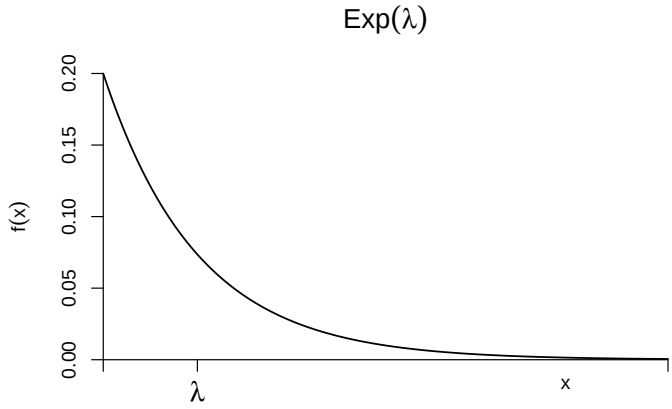
6.5.9 Banaketa normaletik eratorritako banaketak

Jarraian dituzun hiru banaketak (Khi-karratua, Studenten t eta Fisher-Snedecoren F banaketak) banaketa normaletik eratorritakoak dira. Horien erabilera inferentzia-atalean ikusi eta landuko dugu.

6.5.9.1 Khi-karratuaren banaketa

n askatasun-gradudun χ^2 banaketa (ikus 6.3 irudia).

$X \sim \chi^2_n$

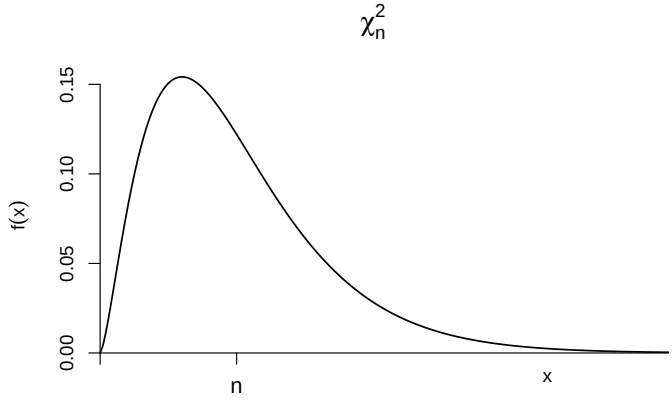


6.2 irudia. Banaketa esponentziala.

Ezaugarriak:

$$E\ X = n \quad \text{eta} \quad VAR(X) = 2n.$$

Eskuin-alboratua da.



6.3 irudia. Khi-karratuaren banaketa.

6.5.9.2 Studenten t banaketa

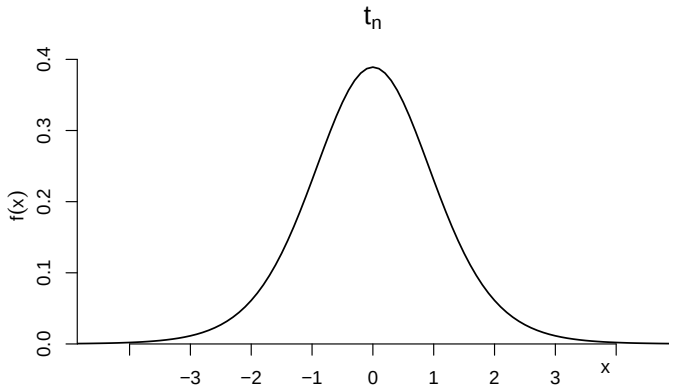
n askatasun-gradudun t banaketa (ikus 6.4 irudia).

$X \sim t_n$

Ezaugarriak:

$$E\ X = 0 \quad \text{eta} \quad VAR(X) = \frac{n}{n-2}.$$

Simetrikoa da.



6.4 irudia. Studenten t banaketa.

6.5.9.3 Fisher-Snedecor-en banaketa

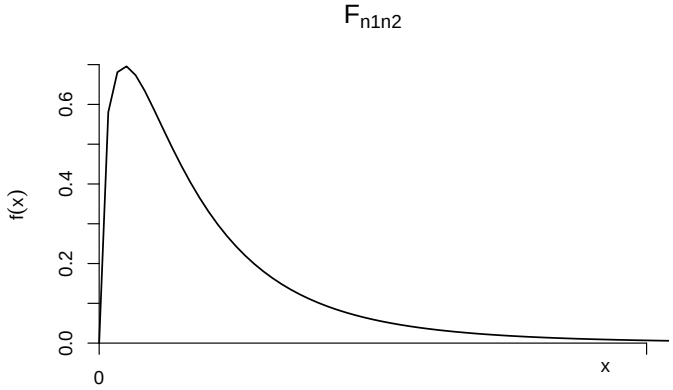
n_1 eta n_2 askatasun-gradudun F banaketa (ikus 6.5 irudia).

$X \sim F_{n_1\, n_2}$

Ezaugarriak:

$$E\ X = \frac{n_2}{n_2 - 2}$$

Eskuin-alboratua da.



6.5 irudia. Fisher-Snedecor-en banaketa.

6.5.10 Banaketa ezagunak R-n

R programazio-lengoaiak banaketa ezagunekin lan egiteko funtzio oso erabilgarriak eskaintzen ditu. Zehazki, banaketa ezagun bakoitzerako, honako lau funtzio hauek aurkituko ditugu:

- Dentsitate-funtzioa (edo probabilitate-masaren funtzioa), $f(x)$.
- Banaketa-funtzioa, $F(x)$.
- Koantilak kalkulatzeko funtzioa.
- Zorizko laginak sortzeko funtzioa.

Gainera, funtzio horiek izendapen-eredu bera jarraitzen dute banaketa guztietan, eta horrek erabilera oso erraza egiten du.

Adibide batekin hasteko, hartu dezagun banaketa normala. Izan bedi X zorizko aldagaia, $X \sim N(\mu = 10, \sigma^2 = 3^2)$ banaketa jarraitzen duena. Ikus dezagun lau funtzio horien erabilera kasu honetarako.

Dentsitate-funtzioa

R-n, banaketa normalaren dentsitate-funtzioa `dnorm()` funtzioan inplementatuta dago. Izan ere, `dnorm` izenaren hasierako *«d»* hizkia ingeleseko *density* hitzetik dator. Funtzio horrek, $x \in \mathbb{R}$ balio bat emanda, $f(x)$ dentsitate-funtzioaren emaitza itzuliko du.

```
# Adibidea: f(8) kalkulatu N(10, 3^2) banaketan
dnorm(5, mean = 10, sd = 3)
# Emaitza:
# [1] 0.03315905
```

Oharra: `dnorm` funtzioan bi atributu garrantzitsu ditugu:

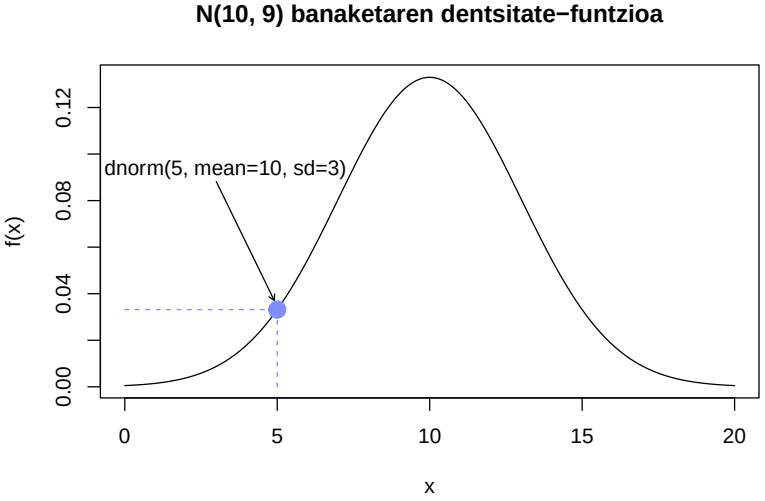
- `mean`: μ parametroari dagokiona.
- `sd`: σ (desbiderapen estandarra) parametroari dagokiona.

Oso garrantzitsua da bi atributu horietan guk erabili nahi ditugun parametroak ezartzea.

Grafikoki, `dnorm`-en emaitza 6.6 irudian adierazita ikus dezakegu.

Ariketa gisa, exekutatu honako kode hau, $N(\mu = 10, \sigma^2 = 3^2)$ banaketaren dentsitate-funtzioa irudikatzeko:

```
x <- seq(0, 20, by = 0.1)
y <- dnorm(x, mean = 10, sd = 3)
plot(x, y, type = "l",
     main = "N(10, 9) banaketaren dentsitate funtzioa",
     ylab="f(x)", xlab="x")
```



6.6 irudia. $N(10, 9)$ banaketaren $f(5)$ balioa, `dnorm` funtzioarekin.

Banaketa-funtzioa

Banaketa normalaren $F(x)$ banaketa-funtzioa `pnorm()` funtzioan inplementatuta dago. Dakigun bezala, funtzio horrek, $x \in \mathbb{R}$ balio bat emanda, $F(x) = P(X \leq x)$ probabilitatea itzuliko digu. Horregatik, `pnorm` izenaren hasierako hizkia (*«p»*) ingeleseko *probability* hitzetik dator.

```
# P(X <= 12) kalkulatzeko:
pnorm(10, mean = 10, sd = 3)
# Emaitza:
# [1] 0.5
```

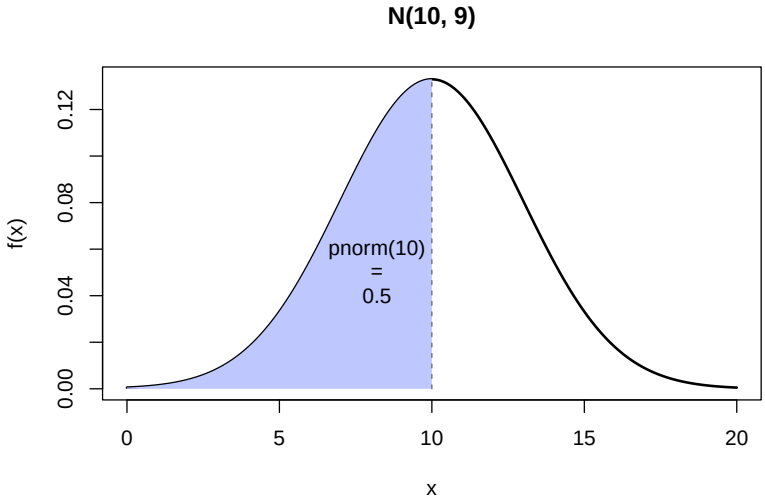
Grafikoki irudikatuta dugu 6.7 irudian.

Koantilen kalkulurako funtzioa

Batzuetan, kontrako arazoa dugu: demagun guk finkatutako edozein $x \in [0, 1]$ baliotarako, $P(X \leq q) = x$ betetzen duen q balioa aurkitu nahi dugula. Hau da, koantil bat. Horrelako kalkuluak egiteko, `qnorm` funtzioa erabil dezakegu. Kasu honetan, hasierako *«q»* hizkiak ingeleseko *quantile* hitzari egiten dio erreferentzia.

Adibidez, X-ren mediana eta 95. pertzentila (P_{95}) kalkulatzeko:

```
qnorm(0.5, mean=10, sd=3)  # P(X <= q) = 0.50 lortzen duen q
                           # q balioa lortzeko
# Emaitza: 10
qnorm(0.95, mean=10, sd=3) # P(X <= q) = 0.95 lortzen duen
                           # q balioa lortzeko
```



6.7 irudia. $N(10, 9)$ banaketan, $F(10) = P(X \leq 10)$ probabilitatea, `pnorm` funtzioarekin.

```
# Emaitza: 14.93456
```

Emaitzak zuzenak direla egiaztatzeko, `pnorm` erabil dezakegu, `qnorm`-en alderantzizko funtzioa baita:

```
pnorm(10, mean=10, sd=3) #P(X <= 10)- = 0.5 dela egiaztatzeko
# Emaitza: 0.5
pnorm(14.93456, mean=10, sd=3) #P(X <= 14.93456)-ren emaitza
                               # 0.95 dela egiaztatzeko
# Emaitza: 0.95
```

Ausazko balioak lagintzeko funtzioa

Azkenik, $N(10, 3^2)$ banaketatik zorizko lagin bat lortu nahi badugu, R-k `rnorm()` funtzioa eskaintzen digu. Kasu honetan, *«r»* letra ingeleseko *random* hitzetik dator. Funtzio horren sarreran, lortu nahi dugun lagin kopurua pasatu beharko diogu (parametro egokiekin batera, aurreko kasuetan bezala). Adibidez:

```
# Ausazko balio bakarra lortzeko:
rnorm(1, mean = 10, sd = 3)
# Emaitza:
# [1] 16.38907  # Erreparatu exekutatzen duzun bakoitzean
                # balio desberdina lortuko duzula

# 5 tamainako lagin bat lortzeko:
rnorm(5, mean = 10, sd = 3)
```

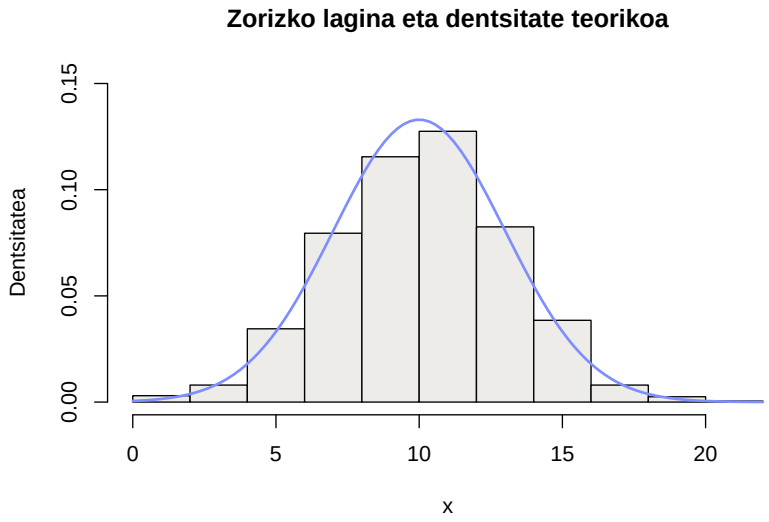
```
# Emaitza:
#[1]  9.058132 13.734718  7.892595  8.232655 10.295490
```

Lortutako balioak $N(10, 3^2)$ banaketatik datozela egiaztatzeko, histograma bat egin dezakegu lagin handi batekin. Histogramaren gainean dentsitate-funtzio teorikoa irudikatuko dugu, argi ikusteko laginak banaketa teorikoaren antz handia lortzen duela (ikus 6.8 irudia).

```
# 1000 balio lagindu N(10,3^2) banaketatik
lagina <- rnorm(1000, mean = 10, sd = 3)

# Histograma, lortutako laginarekin (1000 tamainakoa)
hist(lagina,
     main = "Zorizko lagina eta dentsitate teorikoa",
     ylab="Dentsitatea",  xlab="x", ylim=c(0,0.15), freq=FALSE)

# Ban. normalaren dentsitate teorikoa irudikatzeko kode trinkoa
curve(dnorm(x, mean = 10, sd = 3),
      col = "black", lwd = 2, add = TRUE)
```



6.8 irudia. $N(10, 3^2)$ laginetik zoriz ateratako 1 000 balioen histograma eta banaketaren dentsitate-funtzio teorikoa.

Ohiko banaketen funtzioak: laburpena

Aurreko adibidean, banaketa normalari dagozkion lau funtzioak aurkeztu ditugu. Zorionez, R-n, lau funtzio mota horiek banaketa askotarako inplementatuta daude, eta kasu guztietan berdina da formatua:

- `d<banaketa>`: dentsitate-funtzioa (edo probabilitate-masa funtzioa)
- `p<banaketa>`: banaketa-funtzioa
- `q<banaketa>`: koantilen kalkulua
- `r<banaketa>`: ausazko laginketa

Oharra. Kontutan izan behar dugu banaketa diskretuen kasuan $f(x)$-rekin probabilitate-masa adierazten dela, eta ez dentsitate-funtzioa. Hau da, X zorizko aldagaia diskretua denean, $f(x) = P(X = x)$ probabilitatea da. R-n, ordea, `d<banaketa>` formatua erabiltzen da probabilitate-masa funtziorako ere. Hori nahasgarria izan daiteke, `d` *dentsitate* hitzetik datorrelako.

Oinarrizko banaketei dagozkien funtzioak jaso dira 6.1 taulan, bakoitzaren parametro nagusiekin:

Banaketa teorikoa	Dentsitatea Prob. masa	Banaketa-funtzioa	Koantila	Ausazko laginketa	Parametroak R-n
Binomiala $\mathrm{Bin}(n,p)$	`dbinom`	`pbinom`	`qbinom`	`rbinom`	`size` $= n$ `prob` $= p$
Bin. negatiboa $\mathrm{BinNeg}(r,p)$	`dnbinom`	`pnbinom`	`qnbinom`	`rnbinom`	`size` $= r$ `prob` $= p$
Poisson $\mathcal{P}(\lambda)$	`dpois`	`ppois`	`qpois`	`rpois`	`lambda` $= \lambda$
Uniforme jarraitua $U(a,b)$	`dunif`	`punif`	`qunif`	`runif`	`min` $= a$ `max` $= b$
Normala $N(\mu,\sigma^2)$	`dnorm`	`pnorm`	`qnorm`	`rnorm`	`mean` $= \mu$ `sd` $= \sigma$
Esponentziala $\mathrm{Exp}(\lambda)$	`dexp`	`pexp`	`qexp`	`rexp`	`rate` $= \lambda$
Khi-karratua χ^2_n	`dchisq`	`pchisq`	`qchisq`	`rchisq`	`df` $= n$
Studenten t t_n	`dt`	`pt`	`qt`	`rt`	`df` $= n$
F banaketa F_{n_1,n_2}	`df`	`pf`	`qf`	`rf`	`df1` $= n_1$ `df2` $= n_2$

6.1 taula. Probabilitate-banaketen R-ko funtzio nagusiak

Oharra. Funtzio horien dokumentazioa atzitzeko, `help()` funtzioa erabili. Adibidez: `help(dnorm)`

7

Probabilitate-kalkulua: laginketa, simulazioa eta limitearen teoremak

7.1 Zenbaki handien legea

Zenbaki handien legea. Izan bitez $X_1, \ldots, X_n$ zorizko aldagaiak, denak elkarrekiko askeak eta banaketa berdinarekin (a.b.b.), eta $EX_i = \mu < \infty$. Orduan, $\forall \epsilon > 0$ harturik,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \epsilon\right) \longrightarrow_{n\to\infty} 1.$$

Beste era batera esanda, zorizko aldagaien batezbestekoa populazioko batezbestekorantz hurbiltzen da. Bereziki, jar dezagun $X \sim Bernoulli(p)$ eta errepika dezagun X-ri dagokion esperimentua:

$$\begin{array}{lcl} \text{1. saioan} & \longrightarrow & X_1 \\ \text{2. saioan} & \longrightarrow & X_2 \\ \vdots & & \\ \text{n. saioan} & \longrightarrow & X_n \end{array}$$

Orduan, batezbestekoa hartzen badugu, $\sum_i X_i/n = f_n(X = 1)$, $X = 1$ zein proportziotan gertatu den esango digu, hau da, $X = 1$-en maiztasun erlatiboa adieraziko du. Gainera, $EX = p = P(X = 1)$ denez, zenbaki handien legeak dioenari jarraituz, **maiztasun erlatiboak gertaeraren probabilitatera hurbilduko direla** ondorioztatzen da.

7.2 Limitearen teorema zentrala

Limitearen teorema zentrala. Izan bitez $X_1, \dots, X_n$ zorizko aldagaiak, a.b.b. Demagun $EX_i = \mu$ eta $VAR(X_i) = \sigma^2$ direla. Eraiki dezagun $B_n = \sum_{i=1}^{n} X_i$ batura adierazten duen zorizko aldagaia. Orduan, B_n-ren banaketa asintotikoki normala da. Hau da,

$$B_n \approx N(n\mu, n\sigma^2), \quad n \text{ handia denean.}$$

Limitearen teorema zentrala (LTZ) garrantzitsua da, nahiz eta X_i-ren banaketa nolakoa den ez jakin, hala ere, B_n-ren banaketa normaltzat har dezakegulako n handientzat. n zenbat eta handiagoa izan hurbilpena orduan eta hobea izango da. Erregela orokor gisa, $n \geq 30$ LTZk ematen duen hurbilpena aplika daiteke.



Pentsa ezazu...

LTZren egoerara itzuliz, froga itzazu $E(B_n) = n\mu$ eta $VAR(B_n) = n\sigma^2$ direla.

7.1 adibidea. Zoriz aukeraturiko 100 zenbaki ditugu, 4 hamartarrekin denak. Zenbaki guztien batura egiteko, aurretik zenbaki bakoitza osoetara borobildu da. Kalkula ezazu borobiltze-errorea 5 unitate baino handiagoa izateko probabilitatea.

Ebazpena

Izan bitez,

$$\begin{aligned} Z_i &\equiv \text{«}\textit{Jatorrizko zenbakia}\text{»}\,, & i &= 1, \dots, 100. \\ B_i &\equiv \text{«}\textit{Borobildutako zenbakia}\text{»}\,, & i &= 1, \dots, 100. \end{aligned}$$

Horrela, $Z_i = B_i + e_i$ eta $-0.5 \leq e_i \leq 0.5 \quad i = 1, \dots, 100.$ Gainera, $e_i \sim \mathcal{U}(-0.5, 0.5)$ eta, ondorioz, $E(e_i) = 0$, $VAR(e_i) = 1/12 \quad \forall i.$

Guztizko errorea: $G = \sum_{i=1}^{100} e_i \overset{L.T.Z.}{\approx} N(0, 100/12)$.

Beraz,

$$\begin{aligned} P(|G| > 5) &= P(G < -5) + P(G > 5) \\ &= P(Z < -1.73) + P(Z > 1.73) \\ &= 2P(Z > 1.73) \\ &= 0.0836 \end{aligned}$$

Ondorioak:

Izan bedi $X \sim Bin(n, p)$ (*)

X aldagai binomiala izanik, badakigu berari loturiko esperimentuan, n saiotan lortutako arrakasta kopurua adierazten duela. Hartu ditzagun:

$$\begin{aligned} T_1 &\equiv \textit{«Lehen saioan izandako emaitza»} \\ T_2 &\equiv \textit{«Bigarren saioan izandako emaitza»} \\ &\vdots \\ T_n &\equiv \textit{«n-garren saioan izandako emaitza»} \end{aligned}$$

Honela, $T_i \sim Bernoulli(p)$ $i = 1, \dots, n$ eta askeak dira. Hau da, T_i guztiak askeak eta berdinki banatuak dira, eta, beraz, LTZ aplikatzeko baldintzetan gaude. Ondorioz,

$$\sum_i T_i \approx N(\cdot, \cdot)$$

izango dugu n handirako. Baina, beste alde batetik, $\sum_i T_i = X$ denez,

$$X \approx N(\cdot, \cdot) \quad \text{dugu.} \tag{7.1}$$

Laburbilduz, esperimentu binomialeko saioen kopurua, n, handia denean, banaketa binomialak banaketa normalaren antz handia du.



Pentsa ezazu...

Begira iezaiozu (7.1) ekuazioari. Zein dira normalaren parametroen balioak kasu horretan? Egin itzazu dagozkion kalkuluak.

Guzti horren aplikazio gisa honako erregela hauek erabil ditzakegu:

1. $X \sim Bin(n,p)$ izanik, $np \geq 5$ eta $n(1-p) \geq 5$ badira, orduan $X \approx N(np, np(1-p))$ hurbilpena egokia da.
2. $X \sim Bin(n,p)$ izanik, $n > 50$ eta $p < 0.1$ badira, orduan $X \approx \mathcal{P}(np)$ hurbilpena egokia da.

8

Inferentzia estatistikoa: estimazioa

8.1 Sarrera

Estatistikan inferentzia oso atal garrantzitsua da, eta horri buruz asko dago idatzita (Ikus *Fundamentos de Estadística*, 7. kapitulua [9]). Baina dokumentu honetan, inferentzia zertarako erabil dezakegun ulertzeko, zuzenean adibide batean oinarrituko gara.

Demagun komunikazio-sare batean 0 seinalea bidaltzean 1 seinalea jaso daitekeela, zarata dela eta. Berehala sortzen zaigun galdera hau da: **zein proportziotan gertatzen da hori?**

Galdera horri erantzun egokia eman ahal izateko, egoera hori formalizatzen saiatuko gara.

Hasteko, arazo honetan *populazioa* zein den zehaztu beharko dugu. Kasu honetan, sare horretan bidal daitekeen 0 seinale multzoari dagozkion jasotako seinaleen balioak populazioa osatuko du. Seinale horiek guztiak adierazteko tresna egokia da zorizko aldagaia. Adibidez,

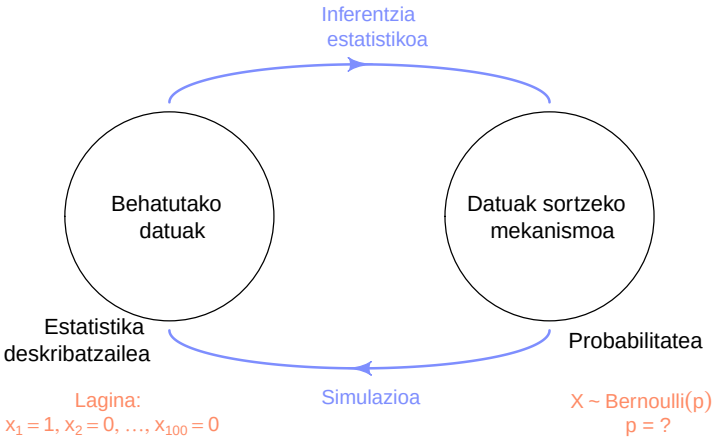
$$X = \begin{cases} 0\,, & \text{akatsik gabe, 0 bidali eta 0 jaso} \\ 1\,, & \text{akatsarekin, 0 bidali baina 1 jaso} \end{cases}$$

eta, beraz, populazioa $X \sim Bernoulli(p)$ moduan adieraz dezakegu.

Behin populazioa modu horretara adierazita, gure helburua p parametro horren balioa nolakoa den jakitea bilakatzen da.

Helburua hori izanik, modurik naturalena komunikazio-sarea martxan jartzea litzateke. Adibidez, esperimentu bat diseinatuko genuke, non 100 aldiz bidaliko baikenuke 0 seinalea baldintza berdinetan eta komunikazio-sarearen beste aldean zer jasotzen den erregistratuko. Demagun esperimentua egin ondoren honelako

lagina behatu dugula: $x_1 = 1, x = 2 = 0, x_3 = 0, \ldots, x_{100} = 0$, guztira $\sum x_i = 15$ behatuz. Orduan, estatistika deskribatzaileko atalean ikusi bezala, batez beste, 15/100=0.15 akats ditugu lagin horretan, eta, gutxi gorabehera, p-ren balioa 0.15 dela asma dezakegu. Baina, baita ere badakigu, esperimentu hori berriro errepikatuko bagenu, ez genituzkeela berriro balio berdinak ikusiko. Hitzez esan duguna, ikasturte hasieran ikusi genuen 8.1 irudian duzu.



8.1 irudia. Estatistika eta Probabilitatea kokatzeko diagrama eta gai hauetan helburu dugun inferentzia-arloa, bi arlo horiek uztartzen dituena.

Inferentzia estatistikoaren lana da gure kezkaren helburu den p-ri buruz informazioa ematea, behatu dugun laginean oinarrituta. Parametro horren benetako balioa ez dugu zehazki inoiz jakiterik izango, baina bai nolako balioak diren «arrazonagarriak». Informazio hori, askotan, bi ikuspunturen arabera planteatuko dugu:

- **Estimazioa.** Parametroaren hurbileko balioak aurkitu eta duten zehaztasuna neurtu. (Gure irakaskuntza-plangintzako 8. gaia).

- **Hipotesi-probak.** Dagoeneko parametroaren inguruko susmo edo hipotesi bat badugu (adibidez, aurretiko komunikazio-sarearen akatsen proportzioa 0.05 da; oraingoarena berdina da?), eta baita datuek hipotesi hori baieztatu ala ezeztatu egiten duten erabakitzeko teknikak ere. (Gure irakaskuntza-plangintzako 9. gaia).

Inferentzian populazioa zorizko aldagai batekin adieraziko dugu. Horretaz gain, zorizko aldagai horren banaketaren itxura orokorra ezagutzen badugu, *parametro-inferentzia* esaten zaio. Hau da, populazioa $X \sim f(x;\theta)$ adieraz dezakegu, eta bertan θ parametroaren inguruko informazioa lortzea izango dugu helburu.

8.2 Populazioa eta lagina

Lehengo adibidera itzuliz, honako egoera hau planteatu dugu:

Populazioa-eredua	**Lagina**
$X \sim \text{Bernoulli}(p)$	$\begin{array}{ccccc} X_1 & X_2 & X_3 & ... & X_{100} \\ \downarrow & \downarrow & \downarrow & & \downarrow \\ 1 & 0 & 0 & ... & 0 \end{array}$

Egoera honetan, p hurbildu/estimatu nahi dugun parametroa da. Hasieratik badakigu laginak ez digula informazio osoa emango, eta, oro har, honako galdera hauek dira erantzun beharrekoak:

1. Nola atera lagina?
2. Nola atera populazioari buruzko informazioa laginetik?
3. Zein konfiantza-maila dute?

Lehenengo galdera erantzuteko, *laginketa* eta *esperimentuen diseinuen* arloan murgildu behar dugu. Irakasgai honetan ez dugu atal hori jorratuko, eta egoera bakoitzean lagin egokia dugula esku artean onartuko dugu.

Beste bi galderak erantzutea izango da gure helburua lantzen ari garen gai honetan. Aurrera egiteko definizio bat behar dugu.

8.1 definizioa. $X_1, X_2, ..., X_n$ zorizko aldagaiek *zorizko lagin bakuna (z.l.b.)* osatzen dute, baldin eta:

i. X_i guztiak elkarrekiko askeak badira.

ii. X_i guztiak berdinki banatuak badira.

Zorizko lagin bakunean $X_i \sim f$ banaketa erakusten badute $(\forall i)$, lagina f banaketadun populaziotik ateratako z.l.b. dela esango dugu. Era berean, populazioa adierazteko f banaketa badugu eta bertatik $X_1, ..., X_n$ z.l.b. atera, $\forall i\ X_i \sim f$ izango da. Adibidez, gure errealitatea edo populazioa adierazteko $X \sim N(\mu, \sigma^2)$ eredua egokia bada, bertatik ateratako z.l.b.-ko elementu bakoitza adierazteko ere $X_i \sim N(\mu, \sigma^2)$ egokia izango da, $i = 1, ..., n$.

8.3 Estimatzaileak eta beren propietateak

8.2 definizioa. Izan bedi $X_1, X_2, ..., X_n$ zorizko lagin bakuna. Aldagai horien gainean osaturiko funtzio errealari *estatistikoa* esaten zaio.

$$T = T(X_1, ..., X_n)$$

8.1 adibidea.

- Batezbestekoa estatistikoa da; izan ere, z.l.b.-ren funtzio erreala da:

$$\bar{X} = \frac{X_1 + ... + X_n}{n}$$

- Gauza bera gertatzen da Estatistika Deskribatzailea atalean ikusitako gainontzeko estatistikoekin.

Kontzeptuekin aurrera egiteko, oso garrantzitsua da erreparatzea estatistikoa **zorizko aldagaia** dela. Hau da, lagin zehatz bat behaturik batezbestekoa $\bar{x} = 7.1$ badugu, «7.1» hori ez da balio finko bakarra, baizik eta lagin zehatz bati dagokiona soilik. Honela, beste lagin bat behatuko bagenu, batez besteko balioa berriro nekez izango litzateke «7.1».



Sakontzeko

Demagun $n = 10$ tamainako z.l.b.-ak aterako ditugula $N(10, 2^2)$ populaziotik. Osa ezazu honako taula hau (erabili, adibidez, Rko `rnorm()` funtzioa):

Lagina l	Behatutako balioak $x_1, x_2, ..., x_{10}$	Behatutako $\bar{X}$-ren balioa $\bar{x}_l$
1		
2		

- Nolako balioak behatu dituzu azken zutabean?
- Uste duzu erraza litzatekeela azken zutabean 6 edo balio txikiagoa suertatzea?

8.3 definizioa. Izan bitez $X_1, X_2, ..., X_n$ zorizko lagin bakuna eta T estatistikoa. T estatistikoaren banaketari *lagin-banaketa* esaten zaio.

Oro har, estatistikoaren lagin-banaketa bi faktoreren menpekoa da: jatorrizko populazioren banaketarena (X_i-en banaketa) eta lagin-tamainarena (n).

8.2 adibidea. Demagun populazioan $X \sim N(\mu, \sigma^2)$ eredua egokia dela. Izan bedi $X_1, X_2, ..., X_n$ bertatik ateratako zorizko lagin bakuna. Nolakoa izango da $\bar{X}$-ren lagin-banaketa?

Ebazpena

Batetik,

$$\bar{X} = \frac{X_1 + ... X_n}{n} \overset{(*)}{\sim} N(\cdot, \cdot)\,, \quad \text{banaketa normalen batura normala delako.}$$

(*) Populazioa ez balitz normala baina $n \geq 30$, limitearen teorema zentralak ondorio berdinera eramango gintuzke.

Orain, gure lana da banaketa normalaren bi parametroak zein izango diren kalkulatzea:

i. Banaketa normalean lehen parametroa itxaropena da, hau da, $E(\bar{X})$ kalkulatu behar dugu. $(\bar{X} \sim N(E(\bar{X})\,, \cdot))$.

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}(X_1 + ... + X_n)\right) \\ &= \frac{1}{n}E\left(X_1 + ... + X_n\right) \\ &= \frac{1}{n}\left(EX_1 + ... + EX_n\right) \\ &= \frac{1}{n}\left(\mu + ... + \mu\right) \\ &= \mu \end{aligned}$$

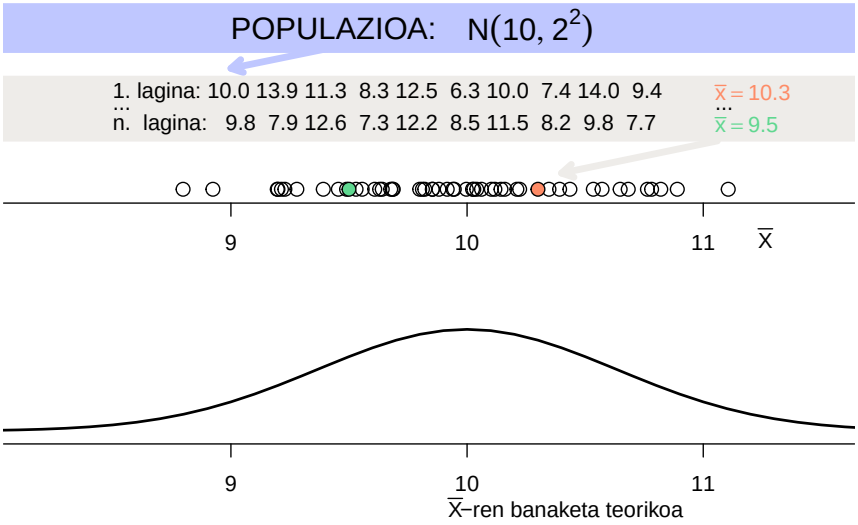
Beraz, $\bar{X} \sim N(\mu, \cdot)$ dela frogatu dugu.

ii. Banaketa normalean bigarren parametroa bariantza da, hau da, $VAR(\bar{X})$ kalkulatu behar dugu. $(\bar{X} \sim N(\cdot\,,\, VAR(\bar{X}))$.

$$\begin{aligned} VAR(\bar{X}) &= VAR\left(\frac{1}{n}(X_1 + ... + X_n)\right) \\ &= \frac{1}{n^2}VAR\left(X_1 + ... + X_n\right) \\ &= \frac{1}{n^2}\left(VAR(X_1) + ... + VAR(X_n)\right) \\ &= \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Oraingoan, $\bar{X} \sim N(\cdot, \frac{\sigma^2}{n})$ dela frogatu dugu.

Dena batera jarriz, $\bar{X}$ estatistikoaren lagin-banaketa $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ dela frogatu dugu (Ikus 8.2 irudia).



8.2 irudia. Populazioa $X \sim N(10, 2^2)$ eran adierazirik, $\bar{X}$-ren *lagin-banaketa.*

Sakontzeko

Demagun aurreko Sakontzeko atalean ezarritakoa dela esku artean dugun egoera. Hau da, $N(10, 2^2)$ banaketatik aterako 10 tamainako lagina:

- Kalkula ezazu kasu horretan $\bar{X}$-ren banaketa.
- Aurreko ataleko emaitzan oinarrituz, kalkula ezazu $P(\bar{X} \leq 6)$, aurreko Sakontzeko azken atalean planteatutako probabilitatea. Zuk uste zenuenarekin bat dator? (Alderatu 8.2 irudiko banaketarekin.)

8.4 definizioa. Demagun θ parametro ezezaguna duen populaziotik ateratako $X_1, X_2, \ldots, X_n$ zorizko lagin bakuna dugula. $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ estatistikoak hartu ditzakeen balioek θ parametroa hurbildu nahi dutenean, $\hat{\theta}$-ri θ parametroaren *puntu-estimatzailea* dela esango diogu.

8.3 adibidea. Estimatzailea vs Estimazioa

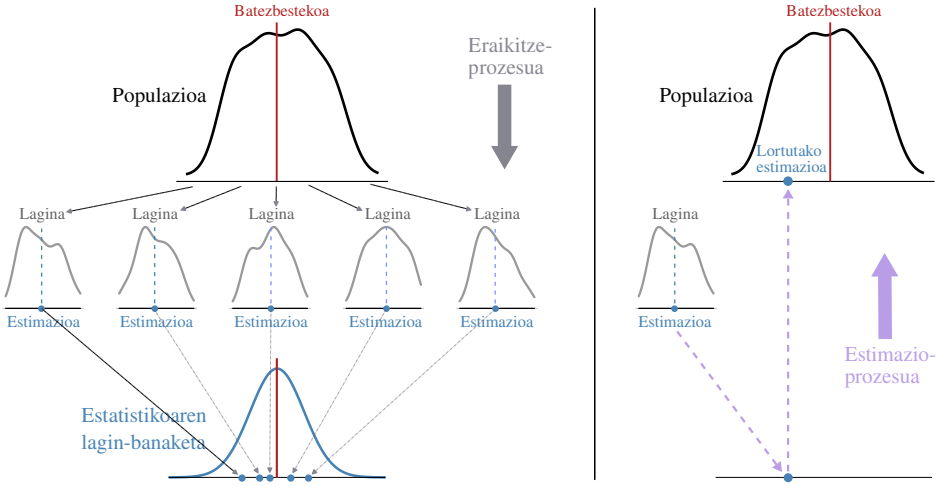
Demagun $X_1, \ldots, X_4$ $N(\mu, \sigma^2)$ populaziotik ateratako z.l.b. dugula eta μ estimatu nahi dugula. Intuizioz, μ-rentzako **estimatzaile** egokia $\hat{\mu} = \bar{X}$ izan daiteke[a].

Baldin eta $x_1 = 3, x_2 = 3, x_3 = 5, x_4 = 6$ z.l.b. behatu badugu, $\hat{\mu} = \bar{X}$ estimatzailearekin μ-ren **puntu-estimazioa** honako hau da:

$$\hat{\mu} = \frac{3+3+5+6}{4} = 4.25$$

[a]Saiatu zaitez arrazonatzen zergatik uste duzun intuizioak hori esan dezakeela.

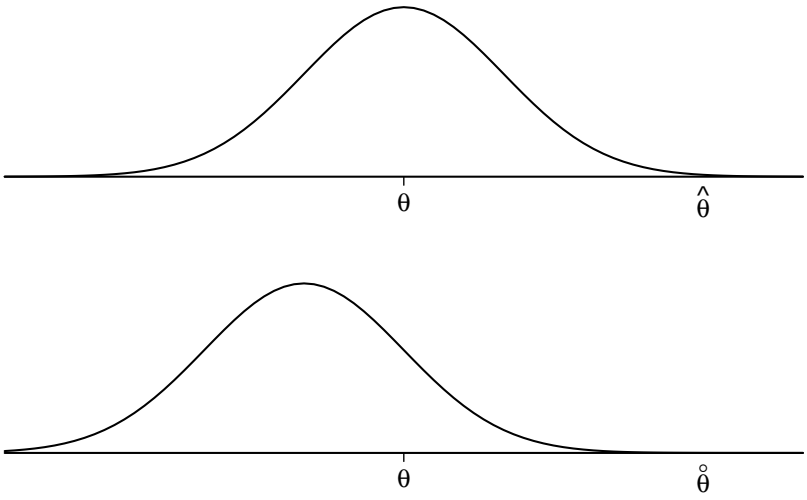
Populazioa, lagin-banaketa eta estimazioa egiterakoan erabiltzen dugun eskema nolakoa den irudikatuta dago 8.3 irudian. Ezkerreko aldean, populazioa eta estatistikoaren lagin-banaketa daude irudikatuta. Estimatzaileak zorizko aldagaiak direla erakusten du. Eskuineko aldean, berriz, estimazio bat egiterakoan ikusten ditugun elementuak soilik daude irudikatuta. Horrela, kasu aplikatu batean estimazioa egiten dugunean jarraitzen dugun prozesua ikus daiteke.



8.3 irudia. Ezkerraldean: populazioa eta estatistikoaren lagin-banaketa; eskuinaldean: estimazio bat.

Jarraian, estimazioan oso garrantzitsua den kontzeptua landu behar dugu. Kontzeptuari dagokion definizioa hobeto ulertzeko adibide bat ikusiko dugu.

8.4 adibidea. Demagun θ parametro ezezaguna dugula eta $\hat{\theta}$ eta $\mathring{\theta}$ bi estimatzaile ditugula. θ estimatzeko, bietako zein den hobea arrazonatzen saiatuko gara. Horretarako, derrigorrez, estimatzaileen lagin-banaketak ezagutu behar ditugu, edo gutxienik beren lagin-banaketez zerbait jakin behar dugu. Demagun azpiko irudikoak direla bi estimatzaileen banaketak.



Lagin-banaketa horiek kontuan hartuta, zein esango zenuke dela estimatzaile hobea θ estimatzeko? Zergatik?

8.5 definizioa. Demagun θ parametro ezezaguna dela eta $\hat{\theta}$ bere estimatzailea. $\hat{\theta}$ estimatzaile *alboragabea* dela esango dugu, baldin eta

$$E(\hat{\theta}) = \theta \quad \text{bada.}$$

8.1 propietatea. *Izan bedi $X_1, \ldots, X_n$ zorizko lagin bakuna, $E\,X_i = \mu$ eta $VAR(X_i) = \sigma^2$ direlarik ($i = 1, \ldots, n$).*

i. $\hat{\mu} = \bar{X}$ estimatzailea alboragabea da μ-rentzat. Hau da,

$$E(\bar{X}) = \mu.$$

ii. $\hat{\sigma}^2 = S_n^2$, lagin-bariantza estimatzaile alboratua da σ^2-rentzat. Zehazki,

$$E(S_n^2) = \frac{n-1}{n}\sigma^2.$$

iii. $\hat{\sigma}^2 = S^2$, kuasibariantza estimatzaile alboragabea da σ^2-rentzat. Hau da,

$$E(S^2) = \sigma^2.$$

Frogapena. Atalez atal egingo ditugu frogapenak.

i. $\hat{\mu}$ estimatzailearen itxaropena kalkulatu behar dugu, eta gure helburu den μ parametroaren berdina dela ikusi.

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_i X_i\right) \\ &= \frac{1}{n}E\left(\sum_i X_i\right) \\ &= \frac{1}{n}\sum_i EX_i \\ &= \frac{1}{n}n\mu \\ &= \mu \quad \square \end{aligned}$$

ii. Frogapena erraz egiteko, gogora dezagun, Y edozein zorizko aldagai izanik, $EY^2 = VAR(Y) + (EY)^2$ dela. Bereziki:

- zorizko lagin bakuneko edozein X_i-tarako:

$$\begin{aligned} E(X_i^2) &= VAR(X_i) + (EX_i)^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Hau da, $E(X_i^2) = \sigma^2 + \mu^2$ (1*).

- $\bar{X}$ aldagaiarentzat:

$$\begin{aligned} E\bar{X}^2 &= VAR(\bar{X}) + (E\bar{X})^2 \\ &= \frac{\sigma^2}{n} + \mu^2 \end{aligned}$$

Hau da, $E(\bar{X}^2) = \sigma^2/n + \mu^2$ (2*).

Horrela,

$$\begin{aligned} E(S_n^2) &= E\left(\frac{1}{n}\sum_i X_i^2 - \left(\bar{X}\right)^2\right) \\ &= \frac{1}{n}\sum_i EX_i^2 - E\left(\bar{X}\right)^2 \\ &= \frac{1}{n}\sum_i (\sigma^2+\mu^2) - \left(\sigma^2/n+\mu^2\right), \text{ (1*) eta (2*) erabilita} \\ &= \sigma^2+\mu^2-\sigma^2/n-\mu^2 \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

Beraz, $E(S_n^2) < \sigma^2$ denez, S_n^2 ez da estimatzaile alboragabea σ^2-rako.□

iii. Kuasibariantzaren eta lagin-bariantzaren arteko erlaziotik:

$$\begin{aligned} E(S^2) &= E\left(\frac{n-1}{n}S_n^2\right) \\ &= \frac{n-1}{n}E(S_n^2)) \\ &= \frac{n-1}{n}\frac{n-1}{n}\sigma^2 \\ &= \sigma^2 \end{aligned}$$

Aldiz, $E(S_n^2) = \sigma^2$ denez, kuasibariantza estimatzaile alboragabea da σ^2-rako.□

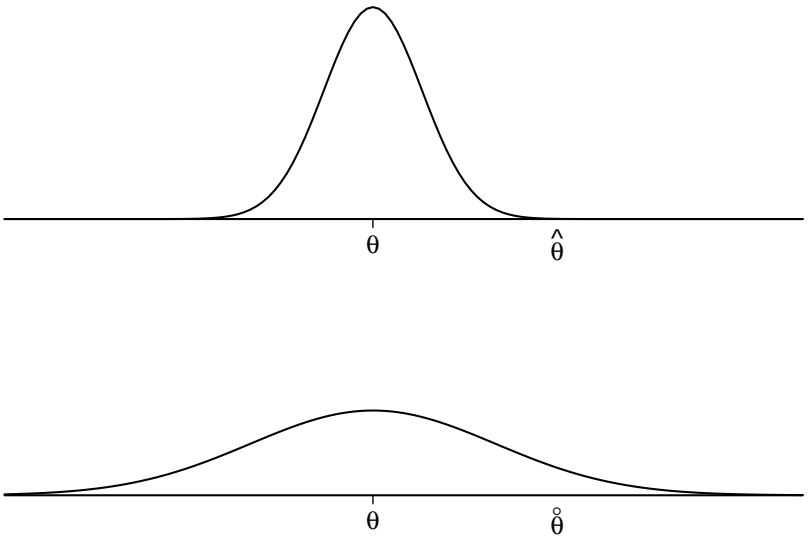


Pentsa ezazu...

Bi estimatzaile desberdin ditugu σ^2 parametroa estimatzeko: s_n^2 eta s^2. Aurreko propietateetan oinarrituz, zein uste duzu dela estimatzaile egokiena? Zergatik?

Estimatzaileen egokitasuna aztertzeko beste propietate garrantzitsu bat eraginkortasuna da. Definizioa hobeto ulertzeko, adibide bat jarriko dugu:

8.5 adibidea. Demagun θ parametro ezezaguna dugula eta $\hat{\theta}$ eta $\mathring{\theta}$ bi estimatzaile ditugula. θ estimatzeko bietako zein den hobea arrazonatzen saiatuko gara. Horretarako, derrigorrez, estimatzaileen lagin-banaketak ezagutu behar ditugu, edo gutxienik beren lagin-banaketez zerbait jakin behar dugu. Demagun azpikoak direla bi estimatzaileen lagin-banaketak.



Lagin-banaketa horiek kontuan hartuta, zein esango zenuke dela estimatzaile hobea θ estimatzeko? Zergatik?

8.6 definizioa. Izan bitez $\hat{\theta}$ eta $\mathring{\theta}$, θ parametro ezezagunaren bi estimatzaile alboragabe. $\hat{\theta}$ estimatzailearen $\mathring{\theta}$ estimatzailearekiko *eraginkortasuna* honela neurtzen da:

$$\frac{VAR(\mathring{\theta})}{VAR(\hat{\theta})}.$$

Baldin eta zatidura hau > 1 bada, $\hat{\theta}$ estimatzailea $\mathring{\theta}$ estimatzailea baino *eraginkorragoa* dela esango dugu.

Sarri, estimatzailearen desbideratze estandarrari *errore estandarra* esaten zaio.



Sakontzeko

Demagun gure populazioa adierazteko eredu egokia dela $N(\mu, \sigma^2)$. Bertatik $n = 2m + 1$ tamainako zorizko lagina atera dugu, n nahikoa handia izanik. μ estimatzeko bi estimatzaile ditugu:

- $\hat{\mu} = \bar{X}$
- $\tilde{\mu} = \text{M}ediana\{X_1, ..., X_{2m+1}\}$

Bi horien artean zein den eraginkorragoa jakin nahi dugu. Horretarako, bi estimatzaileen lagin-banaketak nolakoak diren aztertu behar dugu. Erabili azpiko 8.2 propietatea, zein den eraginkoarragoa erabakitzeko.

8.2 propietatea. *Izan bedi $n = 2m + 1$ tamainako zorizko lagin bakuna, θ mediana eta $f(x)$ dentsitate-funtzioarekin adierazitako populaziotik ateratakoa. Orduan, $f(\theta) \neq 0$ eta deribagarria bada θ-ren ingurune batean, $\hat{\theta} = Mediana_i\{X_i\}$ estimatzailea asintotikoki normala da, θ itxaropena eta $(8mf(\theta)^2)^{-1}$ bariantzarekin.*

8.4 Egoera arruntak

Problema erreal baten aurrean gaudenean, gure helburua parametroen gaineko estimazioa egitea bada, egoera horretarako estimatzaile egokiena zein den topatzea izango da lehen lana. Kontua da, ordea, egoera erreal asko estatistikaren ikuspuntutik berdinak direla eta lan hori egina dagoela dagoeneko arruntak diren egoerak planteatzen dituzten problema errealetan. Hau da, badira egoerak non helbururako estimatzaile egokienen azterketa egina dagoen, eta, beraz, gure lana estimatzaile horiek egoki erabiltzea izango da.

1. X **aldagai kuantitatiboa**

Kasu honetan, gure helburu den errealitatea zorizko aldagai kuantitatibo batez adieraz daiteke. Bi izango dira gure intereseko parametroak. Azpiko taulan daude adierazita estimatu nahi diren parametroak eta dagozkien estimatzaile egokiak.

Parametro interesgarriak	Estimatzaileak
$E(X) = \mu$	$\hat{\mu} = \bar{X}$
$VAR(X) = \sigma^2$	$\hat{\sigma}^2 = S^2$

Estimatzaile horiek erabiliko ditugunez, beharrezkoa da dagozkien lagin-banaketak ezagutzea:

Lagin-banaketak

- Lagin-batezbestekoa, $\bar{X}$.

$$E(\bar{X}) = \mu\,, \quad VAR(\bar{X}) = \frac{\sigma^2}{n}\,, \quad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} \text{ (errore estandarra)}$$

  Gainera, lagina handia ($n \geq 30$) bada: $\bar{X} \approx N(\mu, \sigma^2/n)$.

- Kuasibariantza, S^2.

$$E(S^2) = \sigma^2\,, \quad (*)\ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

  (*)Baldin eta jatorrizko populazioa normala $N(\mu, \sigma^2)$ bada.

8.6 adibidea. Programa baten exekuzio-denborak zenbatekoak diren aztertu nahi da. Horretarako, hainbat datu multzori programa aplikatu, eta exekuzio-denbora hauek behatu dira.

335	568	392	452	479	503	516	398	443	343
316	551	636	363	344	486	518	497	501	310
540	434	415	493	595	405	365	361	433	472
645	332	559	408	455	476	289	561	788	391

Zer esan dezakezu batez besteko exekuzio-denboraren inguruan?

Ebazpena

Interesatzen zaigun egoera laburbil dezakegu honako zorizko aldagai kuantitatibo honekin: $X \equiv$ *«Exekuzio-denbora»*.

Zehazki, interesatzen zaiguna batez besteko exekuzio-denbora da:

$$EX = \mu = ?$$

Balio hori estimatzeko, dagoeneko badakigu erabili behar dugun estimatzailea laginaren batezbestekoa dela, hau da, $\hat{\mu} = \bar{X}$. Beraz, kalkula dezagun dagokion estimazioa gure laginerako:

$$n = 40 \tag{8.1}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{40}\sum_{i=1}^{n} x_i = \frac{1}{40}18368 = \boxed{459.2} \tag{8.2}$$

Horrekin, $EX = \mu$ parametroaren puntu-estimazio bat lortu dugu. Baina, horrez gain, honako galdera hau ebazten saiatuko gara: *beste lagin bat hartuko bagenu, eta berriro laginaren batezbestekoa kalkulatuko bagenu, lortuko genukeen estimazioa oso desberdina izango litzateke? Edo antzekoa?* Galdera hori erantzuteko, errore estandarra erabiliko dugu, hau da, estimatzailearen desbiderapena:

$$SD(\hat{\mu}) = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Izan ere, desbiderapen horrek $\hat{\mu} = \bar{X}$ zorizko aldagaiaren «sakabanapena» kuantifikatzen du, hau da, behaketa desberdinetan (kasu honetan, estimazio desberdinetan) ikusiko dugun aldakortasuna.

Baina, $SD(\hat{\mu})$ kalkulatzerako orduan, segituan arazo batekin topatuko gara: σ parametroa ez dugu ezagutzen. Beraz, ezin dugu errore estandarra kalkulatu. Arazo horrek, ordea, irtenbide erraz bat du. Ikasi dugun bezala, ezagutzen ez ditugun parametroak estima ditzakegu! Nola? Kasu honetan, $\hat{\sigma}^2 = s^2$ bariantza estimatuz laginaren kuasibariantzarekin, ondoren desbiderapena estimatzeko $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{s^2}$ baliokidetzarekin. Era honetan, errore estandarra estimatu ahal izango dugu:

$$\hat{SD}(\hat{\mu}) = \hat{SD}(\bar{X}) = \frac{s}{\sqrt{n}}$$

Aldi berean, s^2 kuasibariantza kalkulatzeko, bide erraz bat da lehenengo s_n^2 lagin-bariantza estimatzea, jarraian $s^2 = \frac{n}{n-1}s_n^2$ baliokidetzarekin kuasibariantza lortzeko. Egin dezagun:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{40}8868474 - (459.2)^2 = 10847.2 \tag{8.3}$$

$$s^2 = \frac{n}{n-1}s_n^2 = \frac{40}{40-1}10847.2 = 11125.3 \tag{8.4}$$

Behin s^2 dugula, s kalkula dezakegu:

$$s = \sqrt{s^2} = 104.15$$

Honekin, $\widehat{SD}(\hat{\mu})$ estimatuko dugu:

$$\widehat{SD}(\hat{\mu}) = \widehat{SD}(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{104.15}{\sqrt{40}} = \boxed{16.7}$$

$\hat{\mu} = 459.2$ puntu-estimazioa eta $\widehat{SD}(\hat{\mu}) = 16.7$ errore estandarra izanda, nekez pentsa dezakegu beste lagin bat hartuko bagenu (berriro puntu-estimazioa kalkulatzeko) emaitza oso desberdina lortuko genukeenik. Esaterako, ez litzateke arraroa izango 440 edo 470 balioak ikustea, baina bai 700 balioa lortzea.

Beraz, ikus daitekeen bezala, $\hat{\mu}$ eta $\widehat{SD}(\hat{\mu})$ balioen baterako interpretazioa oso erabilgarria da, gure intereseko μ balioaren puntu-estimazio bat lortzeaz gain, estimazio horren zehaztasunari buruzko informazioa lortu dezakegulako, gure analisi estatistikoak sendotuz.

2. X **aldagai bitarra**

Kasu honetan, proportzio bat estimatzea dugu helburu, eta azpian aldagai bitar batekin adieraz dezakegu gure egoera edo populazioa. Horrela, $p = P(X = 1)$ izanik, azpiko taulan dugu dagokion estimatzailea.

Parametro interesgarria	Estimatzailea
$E(X) = p$	$\hat{p} = \bar{X}$

Lagin-banaketa

- Lagin-proportzioa, $\hat{p}$.

$$E(\hat{p}) = p\,, \quad VAR(\hat{p}) = \frac{pq}{n}\,, \quad SD(\hat{p}) = \sqrt{\frac{pq}{n}} \text{ (errore estandarra)}$$

Gainera, lagina handia bada ($n\hat{p} \geq 5$ eta $n\hat{q} \geq 5$):

$$\hat{p} \approx N(p, pq/n).$$

8.7 adibidea. Demagun PZU baten aktibitatea neurtu nahi dela, eta, horretarako, $n = 150$ behaketa egin direla. Behaketa bakoitzean (i. behaketa)

$$X_i = \begin{cases} 0 & \text{PZU ez-aktiboa,} \\ 1 & \text{PZU aktiboa,} \end{cases}$$

kodetu da.

Behaketa denak denboran zehar nahiko tarterekin egin dira, horiek elkarrekiko askeak izan zitezen, eta 123 neurketetan PZUa aktibo zegoela ikusi da. Azter ezazu zein proportziotan egoten den aktibo PZUa.

Ebazpena

Gure helburua proportzio bat estimatzea da, zehazki, gure PZUa zein proportziotan dagoen aktibo. Badakigu, horretarako estimatzaile egokia $\hat{p} = \bar{X}$ dela. Badakigu, baita ere, $n = 150$ ale daudela, eta horietatik 123 neurketetan PZUa aktibo zegoela, hau da, $\sum_{i=1}^{n} x_i = 123$.

Beraz, laginean PZUa aktibo egon den proportzioa zuzenean kalkula dezakegu:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{150}123 = 0.82$$

Horrekin, p benetako proportzioaren puntu-estimazio bat lortu dugu. Baina, gehiago egin dezakegu: estimazio horrek zer-nolako zehaztasuna duen ere jakin nahi dugu. Hau da, beste $n = 150$ neurketa berri egingo bagenitu, antzeko balio bat ikusiko genuke? Edo oso desberdina den balio bat ikusi dezakegu? Hori erantzungo digu errore estandarrak:

$$SD(\hat{p}) = \sqrt{\frac{p \cdot q}{n}}$$

Errore estandarra kalkulatzeko, ordea, p-ren benetako balioa jakin beharko genuke. Guk p-ren estimazio bat besterik ez dugunez, errore estandarraren estimazio bat lortuko dugu, $\hat{p}$ erabilita:

$$\widehat{SD}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.82 \cdot (1 - 0.82)}{150}} = 0.031.$$

Dena bateratuz, estima dezakegu 0.82ko proportzio batean ikusten dela aktibo PZUa, eta, horrez gain, estimazio horren zehaztasuna handia dela, % 3.1-eko errore estandarrak adierazten duen bezala. Beste era batera esanda, beste $n = 150$ tamainako lagin bat hartuko bagenu, lortuko genukeen $\hat{p}$ estimazioa ez dela oso desberdina izango ondoriozta dezakegu.

8.5 Konfiantza-tarteak

Orain arte, parametro ezezaguna estimatzeko hurbileko balio bat topatzen genuen zegokion zehaztasunaren neurri batekin. Orain, bi horiek elkartzen saiatuko gara. Ideia orokorraz jabetzeko, hartu dezagun honako kasu hau.

Demagun $X \sim N(\mu, d^2)$ eredua dugula populazioa adierazteko, eta badakigula d^2 **balioa zein den**. Helburua μ batezbestekoaren inguruko informazioa lortzea dugu.

Izan bedi $X_1, \ldots, X_n$ populazio horretatik ateratako z.l.b. ($X_i \sim N(\mu, d^2)$).

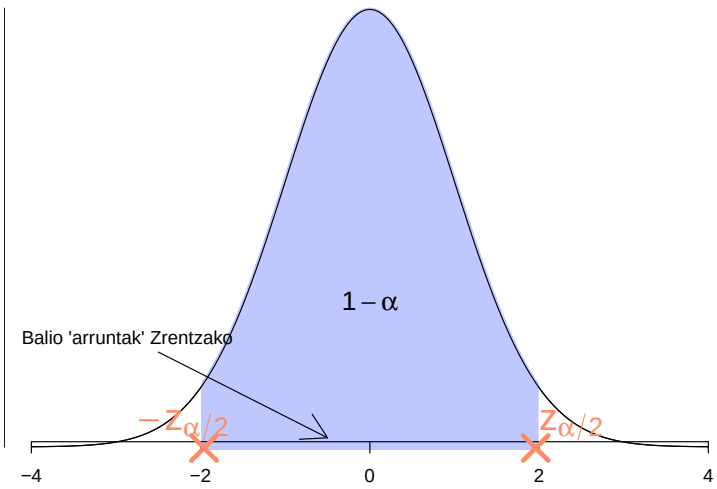
Gure helburua populazio edo ereduko batezbestekoa, μ, estimatzea denez, $\hat{\mu} = \bar{X}$ estimatzailean oinarrituko gara.

Dagoeneko badakigu $\bar{X} \sim N(\mu, d^2/n)$ dela, eta, hortik,

$$\frac{\bar{X} - \mu}{d/\sqrt{n}} \sim N(0, 1).$$

Dei diezaiogun azken horri Z. Hau da, $Z = \frac{\bar{X}-\mu}{d/\sqrt{n}}$.

Zren banaketa ezagutzen dugunez, badakigu zein balio diren espero ditugunak finkatutako probabilitate batekin. Alegia, $P(a < Z < b) = 0.95$ planteatuko bagenu, a eta b balioak zein diren erraz kalkulatuko genuke ($a = -1.96$, $b = 1.96$). Beraz, hori bera planteatuko dugu, baina guk finkatu nahi dugun edozein $1 - \alpha$ probabilitatetarako ($\alpha \in (0, 1)$). Ikus 8.4 irudia.



8.4 irudia. Z estatistikorako balio 'arruntak' $1 - \alpha$ probabilitaterako.

Horrela, $1 - \alpha$ probabilitate baterako balio arruntak $-z_{\alpha/2}$ eta $z_{\alpha/2}$ balioen artean

daude[1].

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1-\alpha \\ P(-z_{\alpha/2} < \frac{\bar{X}-\mu}{d/\sqrt{n}} < z_{\alpha/2}) &= 1-\alpha \\ P(\bar{X} - z_{\alpha/2}\frac{d}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{d}{\sqrt{n}}) &= 1-\alpha \end{aligned}$$

Azken espresioan ikusten dugunez, osa ditzakegu zorizko tarteak non finkatutako probabilitatez helburu den parametroa barneratuko baitute.

8.7 definizioa. Izan bedi $X_1, \ldots, X_n$ θ parametro ezezaguna duen populazio jakin batetik ateratako zorizko lagin bakuna. Demagun baditugula bi estatistiko $T_1(X_1, \ldots, X_n)$ eta $T_2(X_1, \ldots, X_n)$ non:

$$P\left(T_1(X_1, \ldots, X_n) < \theta < T_2(X_1, \ldots, X_n)\right) = 1-\alpha$$

($0 < \alpha < 1$, finkaturiko balioa izanik).

Laginean behatutako balioez egindako $\left(T_1(x_1, \ldots, x_n)\,,\, T_2(x_1, \ldots, x_n)\right)$ tarteari $\%(1-\alpha)100$*eko konfiantza-mailako* θ*-rako konfiantza-tartea* esaten zaio.

$I_\theta^{1-\alpha} = \left(T_1(x_1, \ldots, x_n)\,,\, T_2(x_1, \ldots, x_n)\right)$ eran adieraziko dugu.

Orain ikusi berri dugun adibidearen kasurako, $T_1(x_1, \ldots, x_n) = \bar{x} - z_{\alpha/2}\frac{d}{\sqrt{n}}$ eta $T_2(x_1, \ldots, x_n) = \bar{x} + z_{\alpha/2}\frac{d}{\sqrt{n}}$ dira, eta $I_\theta^{1-\alpha} = (\bar{x} - z_{\alpha/2}\frac{d}{\sqrt{n}}\,,\, \bar{x} + z_{\alpha/2}\frac{d}{\sqrt{n}})$ konfiantza-tartea eskaintzen dute.

[1]Notazioa: z_α esango diogu $P(Z > z_\alpha) = \alpha$ betetzen duen balioari, $Z \sim N(0,1)$ delarik.



Sakontzeko

- Demagun $N(\mu, 4)$ populaziotik $n = 10$ tamainako z.l.b. atera dugula eta honako balio hauek ikusi ditugula:

7.0	5.6	7.1	4.0	8.7	9.3	8.5	8.1	5.4	4.3

  Kalkula ezazu goiko konfiantza-tartearen arabera $I_{\mu}^{0.80}$.
- Aurreko kasu bera hartuz, osa ezazu honako taula hau:

Konfiantza-maila $(1-\alpha)$	Konfiantza-tartea	Tartearen zabalera
0.80		
0.90		
0.95		
0.99		



Pentsa ezazu...

- Konfiantza-tarteak *zorizko* tarteak dira. Zer esan nahi du horrek?
- Demagun lagin batean $I_{\theta}^{0.95} = (3.5, 3.8)$ behatu dugula. Esan dezakegu $P(\mu \in (3.5, 3.75)) = 0.95$ dela?

Lehen ikusi dugun egoerak konfiantza-tarteen kontzeptua ulertzeko balio izan digu, baina, egiari zor, ez da egoera errealista. Hau da, egoera $X \sim N(\mu, d^2)$ ereduarekin adieraz badezakegu, eta gure helburua μ batezbestekoaren inguruko informazioa lortzea izanik, ez da errealista eredu horretan bigarren parametroa ezagutzea. Horregatik, egoera errealagoa planteatuko dugu.

Demagun $X \sim N(\mu, \sigma^2)$ eredua dugula populazioa adierazteko (σ^2 balioa ez dakigu zein den). Helburua da μ batezbestekoaren inguruko informazioa lortzea.

Lehen erabili dugun konfiantza-tartea erabiltzen saiatu gaitezke, baina ez dakigu zein den ereduko bariantzaren balioa. Beste zerbait bilatu behar dugu. Hasieratik hasiko gara berriro.

Izan bedi $X_1, \ldots, X_n$ populazio horretatik ateratako z.l.b. ($X_i \sim N(\mu, \sigma^2)$).

Gure helburua ereduko batezbestekoa, μ estimatzea denez, $\hat{\mu} = \bar{X}$ estimatzailean oinarrituko gara.

Dagoeneko badakigu:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \tag{8.5}$$

baina ezin dugu zuzenean espresio hori erabili konfiantza-tartea eraikitzeko, lehen aipatu bezala (8.5) espresioa σ parametro ezezagunaren menpekoa delako.

Bestetik, estimazioko atalean ikusi dugu σ^2 parametroa estimatzeko $\hat{\sigma}^2 = S^2$ erabil dezakegula. Arrazonagarria dirudi, orduan, honako estatistiko hau eraikitzea:

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}.$$

T estatistikoaren lagin-banaketa zein da, ordea? Hori zehatz-mehatz erantzuteko matematika pixka bat gehiago beharko genuke, eta ideia intuitiboarekin gelditu gara. Gogoratu behar dugu, oro har, T-ren banaketa ez dela N(0,1), baizik eta *Student-en t* banaketa izena duena (ikus 6. gaia). Studenten t banaketak baditu N(0,1)ren antzeko ezaugarriak: itxaropena 0 da, simetrikoa da $y = 0$ ardatzarekiko, baina erabilitako lagin-tamainaren menpekoa da. Gure kasura itzuliz,

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Behin hori jakinda, aurretik erabili dugun arrazonamendu berdinak balio digu kasu honetarako konfiantza-tartea kalkulatzeko. Studenten taula beharko dugu $t_{n-1,\alpha/2}$ balioa kalkulatzeko[2]. Hortik:

$$\begin{aligned}
P(-t_{n-1,\alpha/2} < T < t_{n-1,\alpha/2}) &= 1-\alpha \\
P(-t_{n-1,\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}) &= 1-\alpha \\
P(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) &= 1-\alpha
\end{aligned}$$

Beraz, kasu honetarako, $I_\mu^{1-\alpha} = \left(\bar{x} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\,,\ \bar{x} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$ dugu.

8.5.1 Konfiantza-tarteak lagin-banaketa ezaguna denean

Aurretik ikusitako kasuetan, ezagutzen genuen funtsezko papera jokatzen duen estatistikoaren lagin-banaketa. Lagin-banaketa hori ezagutzen denean, beste edozein kasutarako ere jarraitu beharreko prozedura parekoa da. Jarraian, 8.5 taulan jarri ditugu egoerarik arruntenak laburbilduta.

[2]Notazioa: $t_{n,\alpha}$ esango diogu $P(T > t_{n,\alpha}) = \alpha$ betetzen duen balioari, $T \sim t_n$) delarik.

Parametroa	Baldintzak	**Laburpena** Puntu-estimazioa	Lagin-banaketa	% $(1-\alpha)$-ko K.T.
μ	. σ ezaguna . Pop. normala edo $n > 30$	$\bar{x}$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\left(\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$
μ	. σ ezezaguna . Pop. normala	$\bar{x}$	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$	$\left(\bar{x} \mp t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}}\right)$
μ	. σ ezezaguna . $n > 30$	$\bar{x}$	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \approx N(0,1)$	$\left(\bar{x} \mp z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$
σ^2	. Pop. normala	s^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$	$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}\right)$
p	. Bernoulliren eredua/bin. . $n\hat{p} \geq 5$ eta $n\hat{q} \geq 5$	$\hat{p}$	$\frac{\hat{p}-p}{\sqrt{\hat{p}\hat{q}/n}} \approx N(0,1)$	$\left(\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$

8.5 taula. Egoera arruntenetarako puntu eta tarte bidezko estimatzaileak.

Notazioa:

$$\begin{array}{lcl} z_\alpha & : & P(Z > z_\alpha) = \alpha \\ t_{n,\,\alpha} & : & P(t_n > t_{n,\,\alpha}) = \alpha \\ \chi^2_{n,\,\alpha} & : & P(\chi^2_n > \chi^2_{n,\,\alpha}) = \alpha \end{array}$$

8.8 adibidea. Udaleko Gazteria Sailak agertutako datuen arabera, gazteek gurasoen etxea uzteko adina (X) aldagai normaltzat hartu daiteke. Oraintsu gurasoen etxea utzi duten 16 gazterekin egindako ikerketa batean, 28.1 urteko batezbestekoa lortu da, eta 3ko desbideratze estandarra.

1. Zein da, udal hartako ikerketaren arabera gurasoen etxea uzteko adina batez beste, % 99ko konfiantzaz?
2. Zein da desbideratze estandarraren balioa, gutxi gorabehera, % 99ko konfiantzaz?

1. galderaren ebazpena

Izan bedi $X \equiv$ «Gurasoen etxea uzteko adina». Egoera hori adierazteko eredu egokia $X \sim N(\mu, \sigma^2)$ da. Laginari buruzko informazioa, berriz: $n = 16$, $\bar{x} = 28.1$ eta $s = 3$.

Egoera horretarako konfiantza-tarte egokia $I_\mu^\alpha = \left(\bar{x} \mp t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}}\right)$ da.

Orduan, $1 - \alpha = 0.99 \Rightarrow t_{15,0.005} = 2.95$ aterako dugu taulatik. Dena elkarrekin jarriz:

$$\begin{aligned} I_\mu^{0.99} &= \left(\bar{x} \mp t_{n-1,\,0.005}\frac{s}{\sqrt{n}}\right) \\ &= \left(28.1 \mp 2.95\frac{3}{\sqrt{16}}\right) \\ &= (25.9,\, 30.3) \end{aligned}$$

2. galderaren ebazpena

Kasu honetan, $I_{\sigma^2}^{0.99} = \left(\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}\right)$ da kalkulatu beharreko konfiantza-tartea.

Berriro, $1 - \alpha = 0.99 \Rightarrow \chi_{15,0.995} = 4.6$ eta $\chi_{15,0.005} = 32.8$ aterako ditugu taulatik. Honela:

$$\begin{aligned} I_{\sigma^2}^{0.99} &= \left(\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}\right) \\ &= \left(\frac{15 \cdot 9}{32.8}, \frac{15 \cdot 9}{4.6}\right) \\ &= (4.11,\, 29.35), \end{aligned}$$

eta, beraz: $I_\sigma^{0.99} = (\sqrt{4.11},\, \sqrt{29.35}) = (2.02,\, 5.4)$.

8.6 Bootstrap

Maiz, estimatu nahi den θ parametro horren estimatzailearen lagin-banaketa ezezaguna da. Kasu horietan, *bootstrap* deituriko prozedura erabili daiteke estimatzailearen errore estandarra estimatzeko eta θ-rentzako konfiantza-tartea kalkulatzeko. *Bootstrap* izenez ezagutzen diren ideiak Bradley Efron-ek (1979) aurkeztu zituen lehen aldiz. Populazioko parametro baten inguruko inferentzia egiteko, lagina populazio gisa hartzean oinarritzen da bootstraparen funtsezko ideia. Horrela, n

tamainako laginetik n tamainako lagin berriak sortu behar dira, aleak modu askean aukeratuz. Eskema orokorra honako hau da:

1. Demagun $x_1, \ldots, x_n$ dela behatu dugun zorizko lagin bakuna. $\mathbf{x} = (x_1, \ldots, x_n)$ deituko diogu. Hortik, gure intereseko parametroa estimatu: $s(\mathbf{x})$
2. Bootstrap lagin bat atera: $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$. Horretara, jatorrizko laginetik n ale aukeratu behar dira modu askean; alegia, ale bakoitza behin baino gehiagotan aukeratua izan daiteke. Adibidez, demagun $\mathbf{x} = (x_1, \ldots, x_7)$ behatu dugula; orduan, bootstrap lagin bat hau izan daiteke: $\mathbf{x}^* = (x_5, x_7, x_5, x_4, x_7, x_3, x_1)$. Hor ikusten dugu, adibidez, 5. alea 2 aldiz dagoela eta 2. alea, berriz, ez da inoiz aukeratua izan. Kalkulatu aukeratuta dugun estatistikoa bootstrap laginean: $s(\mathbf{x}^*)$
3. Aurreko atala B aldiz errepikatu. Horrek $s(\mathbf{x}^{*1}), \ldots, s(\mathbf{x}^{*B})$ bootstrap-errepikapen sortuko dizkigu. Horiekin lan eginez burutuko da populazioko parametroaren gaineko inferentzia.

Goiko eskemako ideiari jarraituz, kasuan kasu jarraitu beharreko urratsak desberdinak dira. Oso arlo zabala da hemendik eratortzen dena; guk pare bat kasu bakarrik landuko ditugu.

8.6.1 Errore estandarra estimatzeko

Demagun $\mathbf{x} = (x_1, \ldots, x_n)$ zorizko lagina behatu dugula eta populazioko θ ezezaguna estimatu nahi dugula. Demagun badugula horretarako estimatzaile egokia eta laginean honako estimazio hau neurtzen dugula: $\hat{\theta} = s(\mathbf{x})$. Zein zehaztasun du, ordea, estimazioak? Horretarako, errore estandarra estimatu nahi dugu, eta bootstrapak horretan lagunduko digu. Jarraian duzu horretarako algoritmoa:

Errore estandarrak estimatzeko bootstrap algoritmoa

1. Aukeratu B bootstrap lagin aske, $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$, bakoitza n aleduna. Aleak jatorrizko $\mathbf{x}$ laginetik aukeratuak dira modu askean, hau da, ale bat bakoitzean aukeratua izateko probabilitatea $1/n$ da. *Oharra: askotan nahikoa da B 25-200 tarteko zenbaki bat izatea.*
2. Kalkulatu bootstrap-lagin bakoitzean estatistikoaren balioa:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}), \quad b = 1, 2, \ldots, B$$

3. $\hat{\theta}$-ren errore estandarra honela estima daiteke:

$$\widehat{SD}_B = \sqrt{\sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2/(B-1)}$$

non $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$ baita.

	B				
$\bar{x}$	30	50	100	200	$\sigma/\sqrt{n}$
24.28	0.443	0.423	0.462	0.418	0.447

8.6 taula. Bootstrap bidezko errore estandarren estimazioak batezbestekorako eta benetako errore estandarra (8.9 adibidea).

8.9 adibidea. Hartu dezagun $X \sim N(24, 9)$ populazioa. Demagun ez ditugula ezagutzen populazioko parametroak eta μ, populazioko batezbestekoa ezagutu nahi dugula.

Badakigu inferentzia estatistikoak horretarako bidea irekitzen digula estimazioaren bidez. Horretarako, zorizko laguna behatu behar dugu. Demagun $n = 45$ tamainako zorizko laguna behatu dugula eta batezbestekorako puntuestimazioa 24.28 kalkulatu dugula.

Hurrengo urratsa izango da estimazio horrek nolako zehaztasuna duen neurtzea. Informazio hori estimatzailearen errore estandarrak emango digu. Errore estandarra estimatzeko bide egokia izan daiteke bootstrapak ematen digunа.

8.6 taulan ikusi ditzakezu bootstrap bidez estimatutako errore estandarrak B balio desberdinetarako eta baita errore estandar teorikoa ere. Oso desberdinak dira?

8.6.2 Konfiantza-tarteak kalkulatzeko pertzentilen metodoa

Behin bootstrap-metodoen azpian dagoen mekanismoa barneratuta, erraza da metodo horri jarraitzea.

Konfiantza-tarteak kalkulatzeko pertzentilen bootstrap-algoritmoa

1. Aukeratu B bootstrap-lagin aske, $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$, bakoitza n aleduna. Aleak jatorrizko $\mathbf{x}$ laginetik aukeratuak dira modu askean; hau da, ale bat bakoitzean aukeratua izateko probabilitatea $1/n$ da. *Oharra: B-ren balio handiak behar izaten dira egoera gehienetan ($B \approx 2000$).*

2. Kalkulatu bootstrap-lagin bakoitzean estatistikoaren balioa:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}), \quad b = 1, 2, \ldots, B$$

3. Pertzentilen metodoaren arabera, $1 - \alpha$ konfiantza-tartea honela definitzen da:

$$\left[\hat{\theta}_B^{(\alpha/2)},\ \hat{\theta}_B^{(1-\alpha/2)}\right]$$

non $\hat{\theta}_B^{(\alpha)}$ balioa $\alpha \times 100$ pertzentil enpirikoa baita, B bootstrap-laginetan kalkulatua.

Metodo erraza eta egoera askotan baliagarria da. Dena den, metodo horren zuzenketa/hobekuntza da BC_a deituriko metodoa (*Bias-corrected and accelerated*), eta praktikan azken hori erabiltzea hobesten da.

8.10 adibidea. Berriro ere hartu dezagun $X \sim N(24, 9)$ populazioa. Demagun ez ditugula ezagutzen populazioko parametroak, eta μ populazioko batezbestekoa ezagutu nahi dugula. Kasu honetan, bootstrapa erabiliko dugu μ-rako % 95eko konfiantza-tartea kalkulatzeko, aurreko 8.9 adibideko zorizko lagin bera behatuta ($n = 45$) eta metodo arruntarekin alderatuta. Emaitzak 8.7 taulan daude.

$I_\mu^{0.95}$ bootstrap bidez	(23.43, 25.11)
$I_\mu^{0.95}$ arrunta	(23.44, 25.11)

8.7 taula. μ-rako % 95eko konfiantza-tarteak, bootstrap bidez ($B = 2000$) eta metodo arruntarekin, 8.10 adibideko kasurako. Erreparatu, kasu honetan, parametroaren benetako balioa $\mu = 24$ dela.

9

Inferentzia estatistikoa: hipotesi-probak

9.1 Sarrera

Hipotesi-probak inferentzia estatistikoaren atal garrantzitsua dira. Zer-nolako egoerak ebatzi ditzaketen ikusi eta azpialdean oinarri duten logika zein den ezagutzeko fikziozko egoera bat planteatuko dugu.

Demagun etxe jakin bateko liburu elektronikoaren bateria % 100ean kargatzeko behar den denborak banaketa normalari jarraitzen diola ikusi dela. Zehazki, $X \sim N(\mu, 0.3^2)$. Etxeak dioenez, batez beste 4.5 ordu behar dira bateria guztiz kargatzeko. Zalantzak daude, ordea, eta hori horrela den ala ez argitzeko zorizko lagin bat hartzea erabaki da. Behaturiko datuak azpialdeko taulan bildu dira. Horien arabera, liburu elektronikoa ekoizten duen etxeak arrazoi duela onar al daiteke?

4.8	4.5	4.7	4.7	4.8	4.7	4.8	4.4	4.5	5.1
5.0	4.3	4.3	4.6	5.1	5.0	4.6	5.1	4.5	4.9
4.9	4.5	4.8	4.6	4.7	4.5	4.5	4.3	4.1	4.6

$\Rightarrow n = 30 \quad \sum x_i = 139.9$

Hau da, egoera horretan $X \equiv$ *«Bateria kargatzeko behar den denbora»* izanik, egoera edo populazioa adierazko eredu hau dugu: $X \sim N(\mu, 0.3^2)$.

Liburu elektronikoen etxeak dio $\mu = 4.5$ dela. Itxuraz, ordea, horren inguruko zalantzak daude, hau da, $\mu \neq 4.5$ izan daiteke. Lagineko datuek zein aukeraren

alde egiten dute? Hau da, bi aukera planteatzen ditugu:

$$H_0 : \mu = 4.5$$
$$H_1 : \mu \neq 4.5$$

eta, orain, gure lana bi horietako bat aukeratzea izango da. Alegia, hipotesi-test batean *erabaki bat* hartu behar dugu: H_0 egiazkoa dela onartuko dugu ala H_1 aukeratuko dugu testaren amaieran. Ikus dezagun nola egin dezakegun erabaki batera iristeko.

Gure interesa populazioko batezbestekoan dagoenez, $\bar{X}$-ren lagin-banaketarekin hasiko gara. Dagoeneko badakigu honako hau:

$$\bar{X} \sim N(\mu, \sigma^2/n) = N(\mu, 0.3^2/n) \;\Rightarrow\; Z = \frac{\bar{X} - \mu}{0.3/\sqrt{n}} \sim N(0,1)$$

Zer gertatuko litzaioke Z estatistikoari H_0 egiazkoa balitz? Kasu honetan, erraz erantzun dezakegu.

H_0 egiazkoa balitz:

$$Z = \frac{\bar{X} - 4.5}{0.3/\sqrt{n}} \sim N(0,1) \quad \text{litzateke.}$$

Horrek esan nahi du, H_0 egiazkoa balitz, 0 inguruko balioak espero behar genituzkeela Z-rentzat. Hau da, bai balio oso positiboak edo bai balio oso negatiboak oso arraroak liratekeela H_0 egiazkoa balitz.

Guk lagin bat badugunez, gure laginean gertatu dena eta H_0-pean gertatzea esperoko genukeena konparatuko ditugu erabakia hartzeko. Gure laginean honako hau dugu:

$$n = 30\,, \quad \bar{x} = 139.9/30 \quad \Rightarrow \qquad z_{beha} = \frac{4.663 - 4.5}{0.3/\sqrt{30}}$$
$$\bar{x} = 4.663 \qquad\qquad z_{beha} = 2.98$$

Laginean behatu duguna (z_{beha}) eta H_0 egiazkoa balitz espero duguna zenbateraino bat datozen kalkulatzeko, «behatu duguna bezain arraroa edo arraroagoa» gertatzeko probabilitatea kalkulatuko dugu, betiere H_0-pean:

$$p = 2P(Z > 2.98) = 0.002882$$

Kasu honetan ikusten dugu probabilitate hori oso txikia dela, hau da, datuetan behatu duguna eta H_0-pean espero duguna ez datozela guztiz bat, badagoela nolabaiteko «desadostasuna». Horrela, erabakia hartzeko bi aukera ditugu:

1. H_0 egiazkoa dela onartzen jarraitu, ikusi dugun «desadostasuna» zorizkoa dela onartuz. Hots, zoriz behatu dugun lagina arraroa suertatu dela onartu.
2. Ikusi dugun «desadostasunaren» arrazoia H_0 egiazkoa ez izateari egotzi, hau da, H_0 errefusatu eta H_1 ontzat hartu.

Hipotesi-proben logikan, bi aukera horietatik bigarrena hartzen da. Laginean behatutakoaren eta H_0-pean esperotakoaren arteko desadostasuna beti H_0 egiazkoa ez izateari egotzita. Hasieran planteatu dugun egoerara bueltatuta, liburu elektronikoen etxeak esaten duena ezeztatuko genuke, H_1 onartuz eta, beraz, bateriek kargatzeko behar duten batez besteko denbora ez dela 4.5 ondorioztatuz.

Horiek izanik hipotesi-proben ezaugarri nagusiak, bertan azaldu diren elementuak pixka bat zehatzago aztertu behar ditugu.

9.2 Oinarrizko definizio eta kontzeptuak

9.1 definizioa. Populazioaren gainean ezarritako susmoa *hipotesi estatistikoa* da.

Hipotesi-probetan bi hipotesi kontrajartzen dira:

- *Hipotesi nulua* (H_0).
- *Ordezko hipotesia* (H_1).

Hipotesi-proba burutzeak bi hipotesi horien artean bat aukeratzea esan nahi du, betiere, lagineko datuek erakusten dutena oinarri harturik. Gainera, prozedura honela ezartzen da: H_0 hipotesia egiazkotzat hartzen da, baldin eta datuek ez badute kontrakoa erakusten.

9.1 taulan erakusten den bezala, erabaki okerrak hartu ditzakegu, eta horregatik izango da garrantzitsua erabaki okerrei dagozkien probabilitateak neurtzea.

		H_0 egiazkoa da	H_0 ez da egiazkoa
Erabaki	H_0 onartu	ZUZENA	II motako errorea
posiblea	H_0 errefusatu	I motako errorea	ZUZENA

9.1 taula. Hipotesi-probari dagozkion egoera posibleak

9.2 definizioa. I motako erroreari probaren *adierazgarritasun-maila* (α) esaten zaio.

$$\alpha = P(H_0 \text{ errefusatu } | H_0 \text{ egiazkoa})$$

9.3 definizioa. II motako errorearekin lotuta, probaren *ahalmena* honela definitzen da:

$$P(H_0 \text{ errefusatu } | H_0 \text{ faltsua}) = 1 - P(H_0 \text{ onartu } | H_0 \text{ faltsua})$$

9.4 definizioa. H_0 egiazkotzat harturik, *p-balioa* esaten zaio laginean behatutako emaitza bezain arraroa edo arraroagoa zoriz lortzeko probabilitateari.

Hipotesi-probak eraikitzeko prozedura klasikoak, α adierazgarritasun-maila finkatuz, ahalik eta ahalmen handiena duen hipotesi-proba eraikitzea du helburu. Kasuan kasuko hipotesi-proba egokia topatzea ezinbestekoa izango da, baina edozein delarik erabiliko den hipotesi-proba, honako urrats hauetan oinarrituko da.

9.3 Hipotesi-proba egiteko urratsak

Demagun populazioan ezezaguna den θ parametroaren gainean hipotesi-proba egin nahi dugula.

1. Hipotesiak zehaztu. Oro har, bi mota daude:
   (a) Alde batekoa
      i. Ezkerretikoa $\begin{cases} H_0: & \theta \geq \theta_0 \\ H_1: & \theta < \theta_0 \end{cases}$
      ii. Eskuinetikoa $\begin{cases} H_0: & \theta \leq \theta_0 \\ H_1: & \theta > \theta_0 \end{cases}$

   (b) Alde bikoa $\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$

   Oharra: θ_0 balioa erabiltzaileak kasu bakoitzean finkatuko duen balioa da, eta erreparatu $\theta = \theta_0$ kasua beti hipotesi nuluan barneratzen dela.
2. Adierazgarritasun-maila, α, finkatu.
3. Estatistiko egokia aukeratu, T, eta H_0 **egiazkoa dela suposatuz** laginean estatistikoak hartzen duen balioa kalkulatu: T_{beha}
4. p-balioa kalkulatu.
5. p-balioa eta α adierazgarritasun-maila konparatu, eta horren arabera erabakia hartu:
   - $p \leq \alpha$ bada, H_0 errefusatu. Hau da, H_0 faltsua dela erabaki.
   - $p > \alpha$ bada, H_0 ezin dugu errefusatu. Hau da, H_0 faltsua dela erabakitzeko ez dugu nabaritasunik.

9.1 adibidea. Bi txanpon ditugu, baina orekatuta ez ote dauden zalantza dugu.

- Lehen txanpona 17 aldiz bota ondoren, 5 aldiz atera da aurpegia.
- Bigarren txanpona 100 aldiz bota dugu, baina 29 aldiz atera da aurpegia.

Datu horien arabera, txanponak ez-orekatuak daudela esateko nahikoa nabaritasun ba al dugu?

Ebazpena

Lehen txanpona	**Bigarren txanpona**
$p \equiv$ «*aurpegi-proportzioa*»	$p \equiv$ «*aurpegi-proportzioa*»
$\begin{cases} H_0: & p = 0.5 \\ H_1: & p \neq 0.5 \end{cases}$	$\begin{cases} H_0: & p = 0.5 \\ H_1: & p \neq 0.5 \end{cases}$
Lagina:	Lagina:
$\left.\begin{array}{c} n = 17 \\ \text{Aurp. kop.} = 5(*) \end{array}\right\} \hat{p} = 0.29$	$\left.\begin{array}{c} n = 100 \\ \text{Aurp. kop.} = 29(*) \end{array}\right\} \hat{p} = 0.29$
$E(\hat{p}) = p\,, \quad VAR(\hat{p}) = pq/n\,,$ H_0-pean $\quad \hat{p} \approx N(0.5, 0.5 \cdot 0.5/17)$	$E(\hat{p}) = p\,, \quad VAR(\hat{p}) = pq/n\,,$ H_0-pean $\quad \hat{p} \approx N(0.5, 0.5 \cdot 0.5/100)$
$Z = \dfrac{\hat{p} - 0.5}{\sqrt{0.5^2/17}}$	$Z = \dfrac{\hat{p} - 0.5}{\sqrt{0.5^2/100}}$
$z_{beha} = \frac{0.29-0.5}{\sqrt{0.5^2/17}}$ $= -1.73$	$z_{beha} = \frac{0.29-0.5}{\sqrt{0.5^2/100}}$ $= -4.2$
Eta, beraz:	Eta, beraz:
$p-\text{balioa} = 2P(Z < -1.73)$ $= 2P(Z > 1.73)$ $= 0.0836$	$p-\text{balioa} = 2P(Z < -4.2)$ $= 2P(Z > 4.2)$ $= 3 \cdot 10^{-5}$

(*) *LTZren hurbilpenerako baldintzak betetzen dira.*

Bi txanponen kasuan, intuizioak esango ligukeena eta p-balioak erakusten diguna bat datoz. % 5eko adierazgarritasun-mailaz ($\alpha = 0.05$) ezin dugu esan lehen txapona orekatua ez denik. Aldiz, bigarren txanpona ez dela orekatua esan dezakegu % 5eko adierazgarritasun-mailaz.

9.4 p-balioen inguruan hainbat aipamen

Ikerketa aplikatuaren esparruan, hipotesi-probak aurrera eramaten direnean, azken urteetan gailendu den jokaera honako hau izan da: probaren p-balioari begiratu eta p-balioa < 0.05 baldintza betetzen den ala ez begiratzen da. Betetzen bada, H_0 hipotesia errefusatu eta H_1 hipotesia onartzen da. Normalean, berri-

kuntzari dagokion hipotesia H_1 hipotesian dagoenez, hipotesi hori onartzea lotzen da berrikuntzarekin aurrera egitea. Horrek ekarri du aipaturiko jokaera hori modu itsu eta sistematikoan erabiltzea eta, hori horrela izanik, modu desegokian. Izan ere, p-balioak, beste edozein estatistika-tresna bezala, erabilgarriak dira, baina ezin dira «dogma» moduan erabili, eta, estatistikoki, nekez bereiz daitezke p-balioa=0.051ri eta p-balioa=0.049ri dagozkien probak. Honela, p-balioak modu itsuan erabiltzea ez da egokia, eta hipotesi-probak egitean komeni da honako puntu hauek ere kontuan edukitzea:

- Erakutsi intereseko parametroen estimazioak eta konfiantza-tarteak.
- Erakutsi zenbat hipotesi-froga egin diren.
- Interpretatu p-balioa kontuan izanik lagin-tamaina eta testaren ahalmena.
- Ez erabili p-balioak H_0 hipotesi nulua egiazkoa dela ondorioztatzeko.
- Erabili p-balioak zure ondoriorako beste nabaritasun-iturri bat bezala, baina ez berau bakarra izango balitz bezala.

Begiratu: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond $p < 0.05$, *The American Statistician*, **73**, 1–19. DOI: 10.1080/00031305.2019.1583913

9.5 Lagin bakarrerako hipotesi-proba arruntenak

Estimazioko atalean bezalatsu, badira egoera arrunt batzuk non erabili beharreko estatistikoak eta beren banaketak H_0-pean ezagunak baitira. Jarraian jaso dira egoera horiek, dagozkien estatistikoak eta p-balioen definizioak laburbilduz.

- X **aldagai kuantitatiboa.** $EX = \mu$ **batezbestekorako hipotesi-proba.**

  Baldintzak:

  1. σ ezaguna.
  2. Populazio normala edo $n > 30$.

Hipotesiak	Estatistikoa eta bere banaketa H_0-pean	p-balioa
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$	$Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$p = 2P(Z > \lvert z_{beha} \rvert)$
$\begin{cases} H_0: & \mu \geq \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$		$p = P(Z < z_{beha})$
$\begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$	$z_{beha} = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$p = P(Z > z_{beha})$

- **X aldagai kuantitatiboa. $EX = \mu$ batezbestekorako hipotesi-proba.**

  Baldintzak:

  1. σ ezezaguna.
  2. Populazio normala.

Hipotesiak	Estatistikoa eta bere banaketa H_0-pean	p-balioa
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$	$t = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$	$p = 2P(t_{n-1} > \lvert t_{beha} \rvert)$
$\begin{cases} H_0: & \mu \geq \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$		$p = P(t_{n-1} < t_{beha})$
$\begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$	$t_{beha} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$p = P(t_{n-1} > t_{beha})$

- **X aldagai kuantitatiboa. $EX = \mu$ batezbestekorako hipotesi-proba.**

  Baldintzak:

  1. σ ezezaguna.
  2. $n > 30$.

Hipotesiak	Estatistikoa eta bere banaketa H_0-pean	p-balioa
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$	$Z = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \approx N(0,1)$	$p = 2P(Z > \lvert z_{beha} \rvert)$
$\begin{cases} H_0: & \mu \geq \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$		$p = P(Z < z_{beha})$
$\begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$	$z_{beha} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$p = P(Z > z_{beha})$

- **X aldagai kuantitatiboa. $VAR(X) = \sigma^2$ bariantzarako hipotesi-proba.**

  Baldintzak:

  1. Populazio normala.

Hipotesiak	Estatistikoa eta bere banaketa H_0-pean	p-balioa
$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 \neq \sigma_0^2 \end{cases}$	$\dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$s^2 > \sigma_0^2$ bada, $p = 2P(\chi^2_{n-1} > K_{beha})$ $s^2 < \sigma_0^2$ bada, $p = 2P(\chi^2_{n-1} < K_{beha})$
$\begin{cases} H_0: & \sigma^2 \geq \sigma_0^2 \\ H_1: & \sigma^2 < \sigma_0^2 \end{cases}$		$p = P(\chi^2_{n-1} < K_{beha})$
$\begin{cases} H_0: & \sigma^2 \leq \sigma_0^2 \\ H_1: & \sigma^2 > \sigma_0^2 \end{cases}$	$K_{beha} = \dfrac{(n-1)s^2}{\sigma_0^2}$	$p = P(\chi^2_{n-1} > K_{beha})$

- **X aldagai bitarra. $P(X = 1) = p$ proportziorako hipotesi-proba.**

  Baldintzak:

  1. $n\hat{p}\,, n\hat{q} \geq 5$ $(\hat{q} = 1 - \hat{p})$.

Hipotesiak	Estatistikoa eta bere banaketa H_0-pean	p-balioa
$\begin{cases} H_0: & p = p_0 \\ H_1: & p \neq p_0 \end{cases}$	$Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \approx N(0,1)$	$p = 2P(Z > \lvert z_{beha} \rvert)$
$\begin{cases} H_0: & p \geq p_0 \\ H_1: & p < p_0 \end{cases}$		$p = P(Z < z_{beha})$
$\begin{cases} H_0: & p \leq p_0 \\ H_1: & p > p_0 \end{cases}$	$z_{beha} = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$	$p = P(Z > z_{beha})$

9.6 Hipotesi-proben erabilerarako adibideak

Atal honetan hainbat egoera arrunt planteatu dira. Egoera bakoitzerako hipotesi egokiak planteatzea da helburua, ondoren, dagoeneko laburbildu ditugun hipotesi-proba egokia aplikatzeko eta, azkenik, hortik ondorioztatzen den erabakia hartzeko.

9.2 adibidea. *Full adder* deituriko zirkuituak, bi zenbaki bitar emanik, horien batura eraikitzen du. Batura eraikitzeko behar duen denbora sarreren araberakoa eta zirkuituaren kapazitatearen araberakoa da. Ondorioz, batura eraikitzeko behar den denbora zorizko aldagaia da. Izan bedi T denbora hori, eta demagun bere banaketa normala dela $\sigma^2 = 9$ ns^2 delarik. Horrelako 10 zirkuitu zoriz aukeratu eta batura eraikitzeko, honako denbora hauek behatu dira:

52.4	52.4	53.3	54.7	54.9	55.7	57.0	57.4	57.5	57.6

Batez beste 50 ns baino denbora gehiago behar dutela esan al daiteke % 5eko adierazgarritasun-mailarekin? Adierazi ongi zein diren ezarriko dituzun hipotesiak eta hartuko duzun erabakia.

Ebazpena

Enuntziatuari esker, honako hau dakigu:

$$T \equiv \text{«}\textit{Batura eraikitzeko denbora (ns-tan)}\text{»}, \qquad T \sim N(\mu,\, \sigma^2{=}9)$$

Hau da, populazioa normala da, eta bariantzaren balioa ezaguna da ($\sigma^2 = 9$). Gainera, honako egoera honetan gaudela ikus dezakegu:

- T aldagai kuantitatiboa da.
- $E[T] = \mu$ batezbestekorako hipotesi-test bat burutu nahi dugu.

Hori jakinda, hipotesi-test egokia planteatzea da hurrengo urratsa, hipotesi nulua (H_0) eta ordezkoa (H_1) ezarriz:

$$\begin{cases} H_0 : \mu \leq 50 \\ H_1 : \mu > 50 \end{cases}$$

Behin hipotesiak finkatuta, bi hipotesi horietako bat aukeratzea da helburua, eredu teorikoak eta datuak esaten dutenaren arabera. Horretarako, ikasi dugun hipotesi-testen prozedura jarraituko dugu.

Hasteko, honako bi baldintza hauek betetzen direnez:

- σ ezaguna da,
- populazioa normala da, hau da, $T \sim N(\mu, \sigma^2)$,

orduan, H_0 **egiazkoa dela asumituta**, badakigu honako hau beteko dela:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1), \qquad z_{beha} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Goazen, beraz, z_{beha} estatistikoaren balioa kalkulatzera, gure datuetan behatutako informazioa erabilita:

$$\mu_0 = 50\sigma = \sqrt{\sigma^2} = \sqrt{9} = 3n = 10\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}552.9 = 55.29.$$

$$z_{beha} = \frac{55.29 - 50}{3/\sqrt{10}} = 5.58$$

Orain, laginean behatutako balioa guk espero dugun portaerarekin edo eredu teorikoarekin bat datorren erabaki behar dugu. Test hauetan, hori erabakitzeko tresna nagusia p-balioa da. Eskuinetiko hipotesi-test batekin gaudenez (hau da, $H_1 > \mu_0$ motakoa), orduan p-balioa horrela kalkulatu beharko dugu: $p = P(Z > z_{beha}) = P(Z > 5.58)$. Banaketa normalaren probabilitate-taulatik lortu dezakegu probabilitate hori, edo, zuzenean, R bidez kalkulatu:

Beraz, azken urratsa da p-balioa adierazgarritasun-mailarekin (α) alderatzea. Zehazki, $\alpha = 0.05$ izango da kasu honetan, enuntziatuan adierazten baita % 5eko adierazgarritasun-maila nahi dela. Beraz:

$$p = 1.20 \cdot 10^{-8} \quad < \quad \alpha = 0.05$$

Gure p-balioa α baina txikiagoa denez, gure hipotesi-testaren ondorioa honako hau da: H_0 **errefuxatzea eta** H_1 **onartzea.** Beraz, % 5eko adierazgarritasun-mailarekin, zirkuituak batura eraikitzeko batez beste 50 ns baino denbora gehiago behar duela ondorioztatuko genuke.

9.3 adibidea. Pintura-marka bat ekoizten duen enpresa bateko marketing-departamentua pinturak lehortzen pasatzen duen denbora ezagutzean interesatua dago. Jakin badakite lehortze-denborak 81 min^2 bariantzadun banaketa normala jarraitzen duela. Marketing-departamentuak dioenez, batez besteko lehortze-denbora 75 baino txikiagoa bada, publizitate-kanpaina berezi bat jarriko du martxan. 25 tamainako lagin batean $\bar{x} = 72.3$ minutukoa izan dela behatu da. $\alpha = 0.01$ finkatuz, martxan jarriko al da publizitate-kanpaina berezi hori?

9.4 adibidea. Autoen pneumatikoen marrazkien higadura-denbora aztertu nahi da. Ikertzaileek badakite higadura-denboraren desbideratze estandarra 1 500 ordukoa dela, baina batez besteko higadura-denbora 20 000 ordu baino handiagoa den ikusi nahi dute. Horretarako, 130 pneumatiko zoriz aukeratu dira, eta beren batez besteko higadura-denbora 20 376 ordukoa izan da. Plantea ezazu hipotesi-test egokia, eta esan ezazu % 1eko adierazgarritasun-mailaz ia pneumatikoen batez besteko higadura-denbora 20 000 ordu baino handiagoa den.

9.5 adibidea. Datuak bistaratzeko aplikazio bat garatu dugu, eta aplikazioak memoriara kargatzeko behar duen denbora (X) aztertu nahi dugu. $n = 50$ aldiz kargatu ondoren, 0.91 segundoko batezbestekoa eta 0.19 segundoko desbideratze estandarra behatu dira. Aplikazioa merkaturatu ahal izateko, aplikazioa kargatzeko batez besteko denbora 1 segundo baino gehiago ez dela ikustea da gure helburua. Gure helbururako hipotesi-kontraste egokia plantea ezazu. Zein estatistikotan oinarrituko zara testa burutzeko? Adierazgarritasun-maila $\alpha = 0.05$ finkaturik, kalkula ezazu p-balioa eta hartu ezazu dagokion erabakia.

9.6 adibidea. Zerbitzu informatikoak eskaintzen dituen enpresa bat beste eraikuntza baten bila dabil instalazio berriei leku emateko. Eraikuntza berria oraingo tokitik gertu egotea behar dute langileen eta produktuen higikortasuna baimentzeko. Eguerdian, trafiko gehien dagoenean, badakite toki batetik bestera joateko behar den denborak banaketa normala jarraitzen duela. Printzipioz egokia izan daitekeen eraikuntza bat topatu dute, eta berau aukeratuko dute baldin eta eraikuntza zaharretik berrirako batez besteko denbora 20 minutu baino gehiago ez bada. Erabakia hartzen laguntzeko 10 bidaia egin dituzte, eta hauek dira behaturiko datuak: $\sum_i x_i = 216$, $\sum_i x_i^2 = 4921$. $\alpha = 0.05$ adierazgarritasun-mailarekin, eraikin berri hau aukeratuko al dute?

9.7 adibidea. Aurreko (9.6) adibidearekin jarraituz, eraikuntza zaharretik berrirako denboraren bariantza 25 min^2-koa dela esan al dezakegu, $\alpha = 0.05$ adierazgarritasun-mailaz?

9.8 adibidea. Gozokietan erabiltzen den A substantzia kantitate handiegitan jateak ondorio kaltegarriak eduki ditzake, eta, horregatik, merkatuan dagoen gozokien azterketa bat egin da. Datuen normaltasuna onartuz eta behaturiko lagina honako hau dela jakinik, gozokien A substantzia edukiaren desbideratze estandarra % 4koa dela esan al dezakegu % 5eko adierazgarritasun-mailaz?

8.4	11.9	10.9	10.6	4.6	4.9	10.7	18.5	11.3	10.7	10.6	7.4

9.9 adibidea. Pakete eta fardelak premiaz entregatzen dituen enpresa batek esaten du goizeko 9etarako jasotzen dituen fardelen % 90a gutxienez eguerdiko 12etarako entregatuta duela (hiri barruan egin beharrekoetara mugatuz). 225 fardeletatik % 80a bakarrik entregatu badu garaiz, $\alpha = 0.01$ adierazgarritasun-mailaz esan al daiteke enpresak bere hitza betetzen duela?

9.7 Bi laginetarako hipotesi-proba arruntenak

Orain arte, lagin batean bakarrik erreparatu dugu, eta konfiantza-tarte eta hipotesi-probak parametro bakar baten gainean egin ditugu (μ, σ^2 edo p). Orain gure lana ikusitako metodo horiek zabaltzea izango da, bi populazio desberdinetako parametroak konparatu ahal izateko. Erabili beharreko estatistikoak, noski, desberdinak dira, baina azpian dagoen oinarria eta arrazonamendua berdinak direla ohartuko zara.

1 populazioan	2 populazioan
$X_1 \sim f_1$	$X_2 \sim f_2$
$EX_1 = \mu_1$	$EX_2 = \mu_2$
$VAR(X_1) = \sigma_1^2$	$VAR(X_2) = \sigma_2^2$

Populazio bakoitzak bere ezaugarriak izanik, bakoitzetik n_i tamainako zorizko lagin bakuna ($i = 1, 2$) atera dugun kasuan kokatuko gara.

- **X_1 eta X_2 aldagai kuantitatiboak. $EX_1 = \mu_1$ eta $EX_2 = \mu_2$ batez-bestekoen konparaketarako hipotesi-proba.**

  Baldintzak:

  1. σ_1 eta σ_2 ezezagunak baina berdinak ($\sigma_1 = \sigma_2$).
  2. Populazioak normalak.
  3. Bi laginak elkarrekiko askeak.

Hipotesiak	Estatistikoa eta bere banaketa H_0-pean	p-balioa
$\begin{cases} H_0: & \mu_1-\mu_2=0 \\ H_1: & \mu_1-\mu_2\neq 0 \end{cases}$	$t \overset{(*)}{=} \dfrac{\bar{X}_1-\bar{X}_2}{S_p\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$	$p=2P(t_{n_1+n_2-2}>\lvert t_{beha}\rvert)$
$\begin{cases} H_0: & \mu_1-\mu_2\leq 0 \\ H_1: & \mu_1-\mu_2> 0 \end{cases}$		$p=P(t_{n_1+n_2-2}<t_{beha})$
$\begin{cases} H_0: & \mu_1-\mu_2\geq 0 \\ H_1: & \mu_1-\mu_2< 0 \end{cases}$	$t_{beha}=\dfrac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{1/n_1+1/n_2}}$	$p=P(t_{n_1+n_2-2}>t_{beha})$

$(*)$ $S_p^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

Oharra: test honek $t-test$ izena hartzen du.

- X_1 **eta** X_2 **aldagai bitarrak.** $P(X_1=1)=p_1$ **eta** $P(X_2=1)=p_2$ **proportzioen konparaketarako hipotesi-proba.**

  Baldintzak:

  1. $n_i\hat{p}_i\,, n_i\hat{q}_i\geq 5$, $i=1,2$ $(\hat{q}_i=1-\hat{p}_i)$.
  2. Bi laginak elkarrekiko askeak.

  Izan bitez $p_0=p_1-p_2$ eta $\hat{p}_0=\dfrac{n_1}{n_1+n_2}\hat{p}_1+\dfrac{n_2}{n_1+n_2}\hat{p}_2$.

Hipotesiak	Estatistikoa eta bere banaketa H_0-pean	p-balioa
$\begin{cases} H_0: & p_1-p_2=0 \\ H_1: & p_1-p_2\neq 0 \end{cases}$	$Z=\dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_0\hat{q}_0(1/n_1+1/n_2)}}\approx N(0,1)$	$p=2P(Z>\lvert z_{beha}\rvert)$
$\begin{cases} H_0: & p_1-p_2\geq 0 \\ H_1: & p_1-p_2< 0 \end{cases}$		$p=P(Z<z_{beha})$
$\begin{cases} H_0: & p_1-p_2\leq 0 \\ H_1: & p_1-p_2> 0 \end{cases}$	$z_{beha}=\dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_0\hat{q}_0(1/n_1+1/n_2)}}$	$p=P(Z>z_{beha})$

9.8 Bi laginetarako hipotesi-proben erabilerarako adibideak

Atal honetan, bi laginetako egoerak jasotzen dituzten pare bat adibide jaso dira.

9.10 adibidea. Informatikarien soldatak aztertu nahi dira; bereziki, emakumeen eta gizonezkoen bazte besteko soldatak berdinak diren ala ez aztertu nahi da. Horretarako, $n_1 = 25$ emakumezko eta $n_2 = 30$ gizonezko informatikarien laginetatik urteko soldatak (mila euro) jaso dira. Hauek dira behatutako emaitzak:

Informatikaria	Lagin-tamaina	Batezbestekoa	Desbideratze estandarra
Emakumea	25	25.5	3.6
Gizonezkoa	30	28.7	3.7

Datu horien arabera, emakumezkoen eta gizonezkoen urteko batez besteko soldatak desberdinak direla esan al daiteke? Adieraz ezazu zein adierazgarritasun-maila erabili duzun.

Egin duzun test hori zuzena izateko, zorizko aldagaien banaketaren gainean balditzaren bat bete beharko litzateke?

Ebazpena

Adibide honetan planteatzen den egoera honela adieraz dezakegu:

Batetik, $X_1 \equiv$«*Emakumezkoen soldata*» eta $X_1 \sim N(\mu_1, \sigma^2)$.

Bestetik, $X_2 \equiv$«*Gizonezkoen soldata*» eta $X_2 \sim N(\mu_2, \sigma^2)$.

Batez besteko soldaten inguruan egin denez galdera, planteatzen diren hipotesiak hauek dira:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Orain, laginetatik jasotako informazioaren arabera, bi hipotesietako bat aukeratu beharko dugu.

Laginetik informazioa jasota eta H_0 **egiazkoa dela onartuta**, honako hau beteko dela dakigu:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2},$$

kasu honetan, 53 askatasun graduko t banaketa (t_{53}), eta kalkulatu behar dugun estatistikoa $t_{beha} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1+1/n_2}}$ da. Erreparatu, horrek bi eredu normalek amankomunean duten bariantza estimatzea eskatzen duela. Kasu honetan:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2+-2} = 13.359 \quad \Rightarrow s_p = 3.655.$$

Goazen, beraz, t_{beha} estatistikoaren balioa kalkulatzera, gure datuetan behatutako informazioa erabilita:

$$t_{beha} = \frac{25.5 - 28.7}{3.655\sqrt{1/25+1/30}} = -3.23$$

Orain, laginean behatutako balioa guk espero dugun portaera edo eredu teorikoarekin bat datorren erabaki behar dugu, eta hori p-balioarekin egingo dugu. Alde biko testa planteatu dugunez, $p = 2P(t_{n_1+n_2-2} > |t_{beha}|) = 2P(t_{53} > 3.23)$. t banaketaren probabilitate-taulatik lortu dezakegu probabilitate hori, edo, zuzenean, R bidez kalkulatu:

Azkenik, p-balioa adierazgarritasun-mailarekin (α) alderatu behar dugu. Kasu honetan, $\alpha = 0.05$ aukeratzen badugu:

$$p = 0.002128209 \quad < \quad \alpha = 0.05$$

Gure p-balioa α baina txikiagoa denez, H_0 **baztertu eta** H_1 **onartuko** dugu. Hau da, emakumezkoen eta gizonezkoen batez besteko soldatak ez direla berdinak ondorioztatzen da.

Azkenik, garrantzitsua da testak egiten ditugunean aztertzea zein balditzatan diren egokiak erabilitako estatistikoak. Kasu honetan, batetik, nabarmena da bi laginak elkarrekiko askeak direla, laginen artean ez dagoelako inolako harremanik. Bestetik, ikusi beharko litzateke egokiak direla eredu normalak soldatak adierazteko eta bi ereduetako bariantzak berdinak direla. Hori horrela ez balitz, bestelako prozedurak bilatu beharko lirateke,

9.11 adibidea. Lengoaia naturalaren azterketa guztiz beharrezkoa da itzultzaile automatikoak eraikitzeko. Zure ikerketa-taldeak dagoeneko eraikita dagoen itzultzaile baten gainean (Z itzultzailea) aldaketa batzuk egin ditu, eta horrela itzultzaile berria (B itzultzailea) eraiki. Bi itzultzaileen eraginkortasuna konparatzeko laginak hartu dira: $n_1 = 350$ esalditatik Z itzultzaileak 210 egoki itzuli ditu, eta $n = 380$ esalditatik B itzultzaileak 240 esaldi egoki itzuli ditu. Datu horiek eskutan, esan dezakegu itzultzaile berria zaharra baino eraginkorragoa dela?

10

Ariketak

Liburuxka honetan jasotako ariketa batzuk egileek asmaturikoak dira, eta beste batzuk erreferentzietan aipatutako lanetatik hartu edota moldatutakoak [1, 2, 9, 10, 12, 14].

1. eta 2. gaiak

ESTATISTIKA DESKRIBATZAILEA. Maiztasun-banaketa eta adierazpen grafikoak.

1. Demagun aztertu nahi dugun populazioa EHUko ikasle guztiek osatzen dutela.
   a) Azaldu itzazu azter daitezkeen 6 ezaugarri.
   b) Azterketa burutzeko zer erabiliko zenuke: errolda edo lagin bat?
   c) Sailka ezazu ezaugarri bakoitza, definituriko moten arabera.
   d) Azaldu itzazu ezaugarri bakoitzaren modalitateak.

2. Honako azterketa hauetan, azaldu itzazu: populazioa, ezaugarria(k), modalitateak eta azterketa mota: errolda edo lagina, aldagai bakarrekoa edo anitzekoa.
   a) Makina batek hilabetean ekoitzitako piezen artean, 1 000 pieza aukeratzen ditugu eta beren diametroa neurtzen dugu.
   b) Ekoizpenaren hasieratik X markako ordenagailuen salmenta kopurua urteko.
   c) 500 aleko paketearen orri bakoitzaren lodiera, mm-tan.
   d) Urtean zehar, kirol-klub bateko jokalari bakoitzak izan duen lesio kopurua.
   e) Laborategiko esperimentu batean, 10 metroko kobre-hari batean korronte elektrikoa pasatzeko behar den denbora 100 zatitan neurtzen da

(segundotan).

f) Streaming plataforma baten jarduera hasi zenetik, bezeroak urtean gehien ikusitako film-generoa.

g) Erostetxe handi baten bezeroen artean, egun batean honako hau galdetu da: adina, sexua, anai-arreba kopurua, jaiotze-lekua, lanbidea, pisua kg-tan eta altuera cm-tan.

3. Informatika Fakultatean, 1go mailako ikasleei egindako adimen-testen emaitzak honako hauek dira:

Emaitzak	Ikasle kopurua
50-54	1
55-59	2
60-64	11
65-69	10
70-74	12
75-79	21
80-84	6
85-89	9
90-94	4
95-99	4

Ondo eratuta al daude tarteak? (Galdera ez da tarte kopuruaren ingurukoa, baizik eta ezaugarriaren modalitateek bete behar dituzten baldintzen ingurukoa).

Konpon ezazu taula, eta osa ezazu taula estatistikoa.

4. Airearen kutsadura aztertzeko, 57 hiritan aztertu zen esekidura-materiaren partikula-edukina (mikrogramo/m^3-tan), eta honako emaitza hauek lortu ziren:

68	63	42	27	30	36	28	32	79	27	22	23
65	43	25	74	51	36	42	28	31	28	25	45
12	32	49	38	42	27	31	50	38	21	16	24
22	43	27	49	28	23	19	46	30	43	49	12
24	25	44	12	57	51	69	47	23			

Eratu ezazu azterketa honetarako taula estatistiko egokia, eta baita datu horiek erakusteko adierazpen grafiko egoki bat ere.

5. Osatu honako taula hau. Adierazi modu formalean gelaxketako elementuak nola kalkulatu dituzun:

x_i	n_i	f_i	N_i
1	4		
2	4		
3		0.16	
4	7	0.14	
5	5		28
6			38
7			45
8			

6. Datu-base zerbitzari batek 10 milisegundoko jaso dituen eskaerak honako taula honetan bildu dira.

Eskaera kopurua	0	1	2	3
Maiztasuna	36	26	11	2

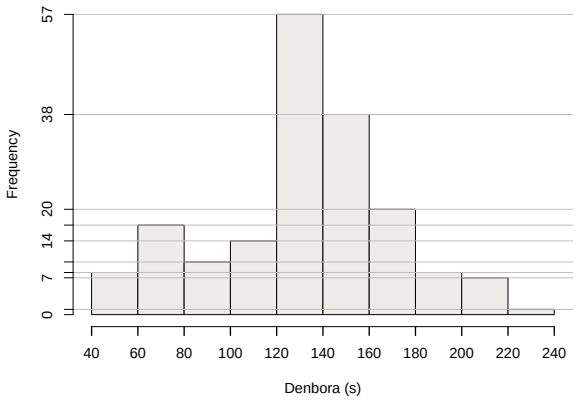
Osa ezazu maiztasun erlatiboen taula, eta marraztu dagokion barra-diagrama.

7. Honako taula honek, lantegi batek manufakturaturiko 60 bolen diametroak (cm) adierazten dizkigu (txikienetik handienera ordenatuta):

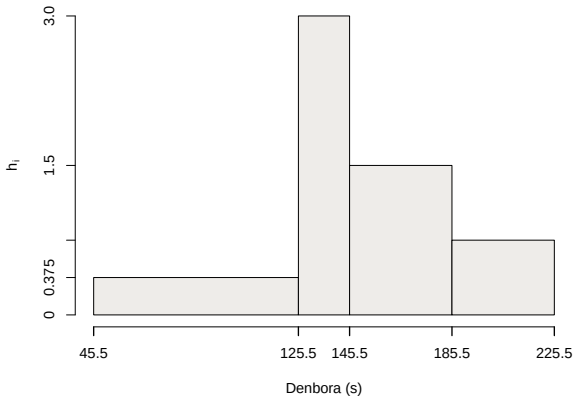
1.714	1.714	1.715	1.717	1.718	1.719
1.720	1.720	1.720	1.721	1.721	1.722
1.722	1.722	1.722	1.723	1.724	1.725
1.725	1.725	1.727	1.728	1.728	1.728
1.728	1.728	1.728	1.729	1.729	1.729
1.730	1.731	1.731	1.733	1.733	1.733
1.733	1.734	1.735	1.735	1.736	1.736
1.736	1.737	1.737	1.738	1.739	1.739
1.740	1.740	1.740	1.740	1.741	1.744
1.746	1.749	1.749	1.750	1.755	1.763

Eraiki ezazu histograma.

8. Software-enpresa batek jasotzen dituen kodeen exekuzio-denborak (milisegundo) honako histograma honetan adierazi dira. Eraiki ezazu dagokion maiztasun-taula.



Demagun aldaketa batzuen ondoren behatutako exekuzio-denborak adierazi direla honako histograma honetan. Eraiki ezazu dagokion maiztasun-taula ($n = 180$).



9. Hilabetean zehar, gaixotasun kronikoetarako ospitale batean onarturiko gaixoen adinak (urtetan) aztertu dira, eta honako datu hauek lortu dira:

Adina	Gaixoak	Adina	Gaixoak	Adina	Gaixoak	Adina	Gaixoak	Adina	Gaixoak
1	2	14	2	29	1	49	1	70	1
2	3	15	1	30	1	51	3	71	1
3	1	16	1	32	2	52	3	72	1
5	2	17	1	33	4	53	4	74	1
6	3	18	1	36	1	54	2	76	1
7	5	21	1	37	1	56	1	77	1
8	1	22	3	39	1	58	2	81	2
9	2	24	2	40	1	60	1	85	4
10	3	25	1	42	2	61	2	87	3
11	1	27	2	46	2	63	2	88	1
12	1	28	2	48	3	69	2	89	1

a) Eraiki itzazu taula estatistikoa eta histograma, datuak honako tarte

hauetan sailkatuz: 1-15, 15-30, 30-45, 45-60, 60-75, 75-90.

b) Egin ezazu lan berdina, baina datuak beste era honetara sailkatuz: 1-5, 5-15, 15-30, 30-45, 45-60, 60-90.

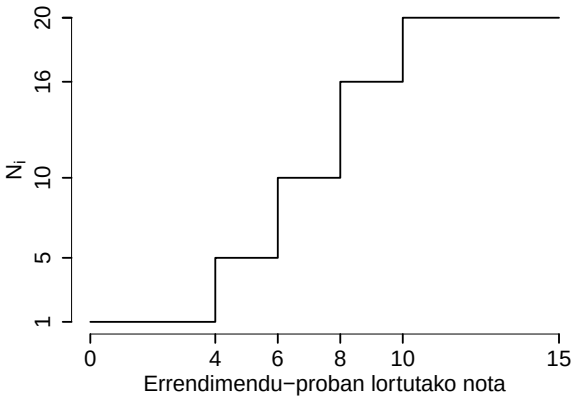
c) Konparatu itzazu emaitzak.

10. Proba egokien bidez, 20 ikasleren adimen-koefizienteak neurtu dira, eta lortutako emaitzak zabalera ezberdinetako sei tartetan banatuta daude. Zabalerak hauxek dira: $a_1 = 12$, $a_2 = 12$, $a_3 = 4$, $a_4 = 4$, $a_5 = 12$ eta $a_6 = 20$. Tarteen maiztasun erlatibo metatuak $F_1 = 0.15$, $F_2 = 0.15$, $F_3 = 0.55$, $F_4 = 0.8$, $F_5 = 0.95$ eta $F_6 = 1.00$ dira. Honako hau eskatzen da:

    a) Lehenengo tartearen beheko muga 70 dela jakinik, eratu ezazu maiztasun-taula.

    b) Irudika ezazu histograma.

    c) Ikasle talde berberari errendimendu-proba bat egin zaio. Eratu ezazu honako adierazpide grafiko honetan agertzen diren emaitzen taula estatistikoa:



11. Programaren exekuzio-denboraren honako behaketa hauek ditugu:

0.68	0.84	1.00	1.08
0.78	0.94	1.01	1.14
0.83	0.95	1.01	1.16
0.83	0.98	1.05	1.19
0.83	0.98	1.07	1.20

Eraiki ezazu puntu-diagrama.

12. Corpus[1] bateko hitzen agerraldien maiztasunak bildu dira honako taula honetan:

[1] Hizkuntzaren prozesamenduan, corpus edo testu-corpusa nolabait egituratutako testu sorta handi bat da. Ikusi wikipedian.

$[l_i, l_{i+1})$	n_i	f_i
$[1, 3)$	13096	0.326
$[3, 6)$	12034	0.300
$[6, 10)$	7370	0.184
$[10, 15)$	5494	0.137
$[15, 50)$	2108	0.053

Eraiki ezazu histograma. Zer erakusten dizu?

3. gaia

ESTATISTIKA DESKRIBATZAILEA. Joera zentrala, sakabanapena eta posizio estatistikoak.

1. Gela bateko 110 ikasleren altuerak neurtu ondoren, honako taula honetan dauden datuak lortu dira:

Altuera	Ikasle kopurua
$[1.55 - 1.60)$	18
$[1.60 - 1.70)$	31
$[1.70 - 1.80)$	24
$[1.80 - 1.90)$	20
$[1.90 - 2.00)$	17

   Kalkula itzazu:

   a) 21. eta 87. pertzentilak, eta 3. eta 9. dezilak.
   b) Kontsideratuko dugu 3. pertzentiletik beherako altuera dutenak «txikiak» direla. Zein da horiek hartu dezaketen altuera maximoa?
   c) Kontsideratuko dugu 82. pertzentiletik gorako altuera dutenak «altuak» direla. Zein da horiek hartu dezaketen altuera minimoa?
   d) Zein pertzentiletan egongo da 1.87 metroko altuera duen gazte bat?

2. Honako zerrenda honetan 35 egunetan zehar enpresa batean jasotako iruditxartel kopuruak adierazi dira:

200	300	100	100	0	200	0
0	200	0	0	400	0	0
400	0	1000	300	0	0	0
0	0	0	100	0	300	0
200	200	0	0	100	300	0

a) Kalkula itzazu banaketa honentzat ezagutzen dituzun joera zentraleko estatistikoak.
b) Kalkula itzazu banaketa honentzat ezagutzen dituzun sakabanapen estatistikoak.
c) Komentatu lortu dituzun emaitzak.

R erabiliz garatzeko:

```
# Sartu behaketak objektu batean.
# Lehen 5 elementuak BAKARRIK sartuko ditugu.
# Zuk osatu gainontzekoekin
x <- c(200, 300, 100, 100, 0)
# n, lagin-tamaina 35 dela ziurtatu
n <- length(x)
n
# Batezbestekoa eskuz kalkulatzeko:
# Balio guztien batura...
sum(x)
# ... zati n
sum(x)/n
# Mediana kalkulatzeko
# Zenbatgarren elementuak zehazten du erdia? n bakoitia denez...
erdikoi <- ceiling(n/2)
#... ordenatu x balioak...
xord <- sort(x)
#... eta hartu erdiko elementuari dagokion balioa
xord[erdikoi]
# Sakabanapen estatistikoen kalkulurako, karratuen baturak
# kalkulatu beharko dituzu
# Zer egiten du x^2 aginduak?
x^2
# Karratuen batura:
sum(x^2)
# Berdinak dira sum(x^2) eta sum(x)^2?
# Idatzi bakoitzari dagokion espresio matematikoa
```

3. Demagun batezbesteko aritmetikoa eta bariantza 7 eta 1 000, hurrenez hurren, dituen maiztasun-banaketa dugula. Batezbestekoa banaketa horren adierazgarria al da?

4. Aldagai estatistiko batek x_i balioak hartzen ditu, dagozkien maiztasun absolutuak $n_i, i = 1, ..., k$ izanik; eta beste aldagai batek $y_i, i = 1, ..., k$, balioak hartzen ditu, dagozkien maiztasun absolutuak berdinak direlarik. Froga itzazu batezbesteko aritmetikoaren honako propietate hauek:

   a) $z_i = x_i + b$ baldin bada, orduan $\bar{z} = \bar{x} + b$ da.
   b) $z_i = ax_i + b$ baldin bada, orduan $\bar{z} = a\bar{x} + b$ da.

c) $z_i = x_i + y_i$ baldin bada, orduan $\bar{z} = \bar{x} + \bar{y}$ da.
d) $z_i = ax_i + by_i$ baldin bada, orduan $\bar{z} = a\bar{x} + b\bar{y}$ da.

5. Informatika Fakultateko 40 ikasleri antsietate-testa egin diete, eta honako datu hauek lortu dira:

Puntuazioa	Ikasle kopurua
$[15, 20)$	2
$[20, 25)$	8
$[25, 30)$	13
$[30, 35)$	7
$[35, 40)$	6
$[40, 45)$	3
$[45, 50)$	1

a) Kalkula ezazu pertsonako batez besteko agresibitatea.
b) Zein puntuaziotik gora egongo dira gelako agresiboenak diren % 25a?
c) Kalkula ezazu desbideratze estandarra.
d) Marraztu itzazu histograman bertan kalkulatutako estatistikoen balioak. Zer esango zenuke banaketaren *forma*ri buruz?

6. Azken urteetan, materialek hegazkinen hainbat zatitan nekeak eraginda sorturiko zartaduren ikerketa sakonak egin dira. Honako datu hauek neurri jakin bateko zartadura sortzeko beharrezko denbora adierazten dute (hega-orduak/10^4):

0.736	0.863	0.865	0.913	0.915	0.937	0.983	1.007
1.011	1.064	1.109	1.132	1.140	1.153	1.253	1.394

a) Kalkula itzazu lagineko batezbestekoaren eta medianaren balioak.
b) Marraztu ezazu datu horiei dagokien puntu-diagrama. Zenbatean gutxitu daiteke laginaren neurri handiena, medianaren balioa aldatu gabe?
c) Kalkula itzazu lagin-bariantza eta desbideratze estandarra. Zein unitatetan neurtzen da estatistiko bakoitza?
d) Egin ezazu kutxa-diagrama bat. Informazioa ardatzean zehar simetrikoki banatua agertzen al da? Balio arrarorik ba al da?

7. $n = 10$ soinu-txarteleko lagina aukeratu, eta horietako bakoitzari kalitate-kontrola pasatu zaio. Kontrola akatsik gabe pasatutako txartela S-z (arrakasta) izendatzen bada eta kalteduna F-z, emaitzak honako hauek dira:

$$S, S, F, S, S, S, F, F, S, S.$$

a) Zein da lagineko arrakasta-proportzioaren balioa, $\widehat{p}$?

b) Ordezka ezazu S bakoitza 1 batekin eta F bakoitza 0 batekin. Kalkula ezazu batezbestekoa, $\bar{x}$, zenbaki bidez kodetutako lagin honentzat.Nola konparatzen dira $\bar{x}$ eta $\widehat{p}$?

c) Suposa ezazu 15 txartel gehiagorekin egiten dela froga. Horietako zenbatek izan behar lukete S, 25 txarteleko lagin osoko $\widehat{p} = 0.8$ izan dadin?

8. Honako balio hauek hainbat eskariren memoria-eskakizunak adierazten dituzte:

$$14.5, 25.6, 52.4, 66.3, 69.3, 69.8, 76.2$$

a) Kalkula itzazu lagineko batezbestekoaren eta medianaren balioak. Bi balio horiek elkarrekin konparatuz gero, zer informazio ematen dute laginari buruz?

b) Kalkula itzazu lagin-bariantza eta kuasibariantza. Zein erlazio dago bi estatistiko horien artean?

9. Aldagai estatistikoaren $x_1, \ldots, x_n$ balioak emanik, kalkula ezazu c-ren zein baliorentzat den $\sum_{i=1}^{n}(x_i - c)^2$ minimoa.

Oharra: c-rekiko deribatua hartu eta minimoa bilatu.

10. Software-enpresa bateko 125 langileen soldatak honela banatzen dira: 7 langilek 2 250 euroko soldata dute, 30 langilek 1 200 eurokoa eta gainontzekoek 1 800 eurokoa. Kalkula ezazu langileen batez besteko soldata.

Oharra: batezbesteko mota horri *batezbesteko haztatua* esaten zaio.

11. Estatistikako azterketa finalean 150 ikasleren batez besteko nota 7.9koa da, eta desbideratze estandarra 2.2koa. Analisi Matematikoko azterketan, berriz, batezbestekoa 6.9koa eta desbideratze estandarra 2.0koa. Zein ikasgaitan da handiagoa:

a) sakabanapen absolutua?

b) sakabanapen erlatiboa?

12. Informatika-ikasketak egiteko ikasleen eskari-igoerak % 9, % 11 eta % 8 izan dira, hurrenez hurren, hiru urtetan zehar. Urte bakoitzean igoera aurreko urtean izandako eskariari aplikatzen zaio. Zein da hiru urte horietan batez beste izandako eskaera-igoera?

Oharra: batezbesteko mota horri *batezbesteko geometrikoa* esaten zaio.

13. Demagun laginean $x_1, \ldots, x_n$ balioak behatu direla.

a) a eta b konstanteak izanik eta $y_i = ax_i + b$, $i = 1, \ldots, n$ aldagai berria osatuz, zein dira $\bar{x}$-ren eta $\bar{y}$-ren arteko erlazioak? Eta $s^2_{n,x}$-ren eta $s^2_{n,y}$-ren artekoak?

b) Erreakzio kimiko bat has dadin beharrezko den laginaren batez besteko tenperatura, Celcius gradutan neurtuta, 87.3 da, eta desbideratze estandarra 1.04. Zein dira laginaren batezbestekoa eta desbideratze estandarra Fahrenheit gradutan? Kontuan izan $^\circ F = 9/5^\circ C + 32$ erlazioa.

14. Demagun $x_1, x_2, \ldots, x_n$ lagina daukagula eta $\bar{x}$, $s^2_{n,x}$ eta $s_{n,x}$ balioak kalkulaturik daudela.

    a) $y_i = x_i - \bar{x}$ bedi, $i = 1, 2, \ldots, n$ denean. Nola konparatzen dira bi aldagaien bariantzak, hau da, $s^2_{n,x}$ eta $s^2_{n,y}$?
    b) $z_i = \frac{(x_i - \bar{x})}{s_{n,x}}$ bedi, $i = 1, 2, \ldots, n$. Zein dira lagin-bariantza eta desbideratze estandarraren balioak z aldagaiarentzat? Transformazio honi x *aldagaia estandarizatzea* esaten zaio.

15. Demagun $x_1, \ldots, x_n$ lagina dugula eta horren batezbestekoa eta lagin-bariantza $\bar{x}_n$ eta s^2_n direla, hurrenez hurren. Demagun, baita ere, lagin honetara x_{n+1} balio berria gehitu dela. Izan bitez $\bar{x}_{n+1}$ eta s^2_{n+1} lagin handituaren batezbestekoa eta lagin-bariantza, hurrenez hurren.

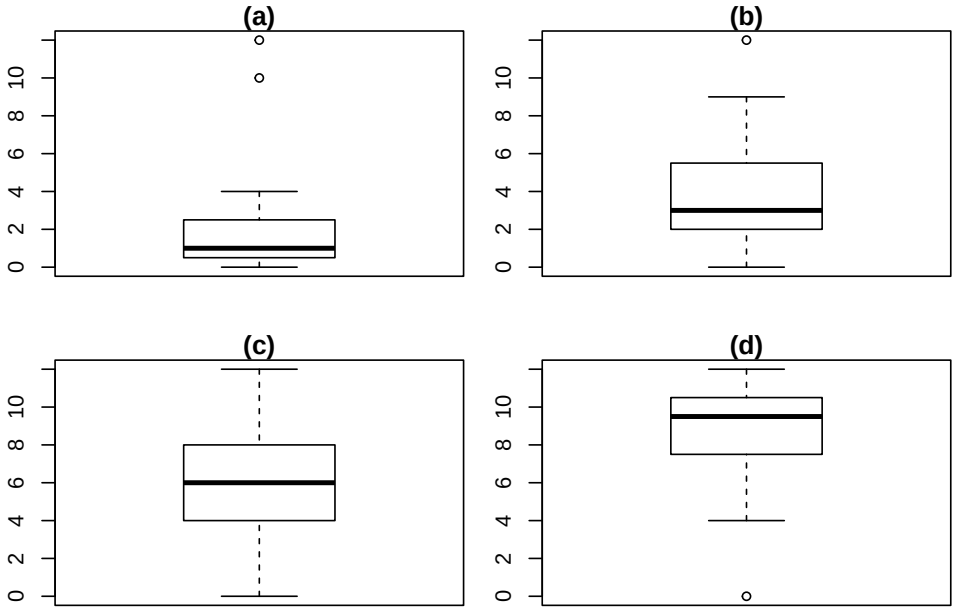
    a) Erakuts ezazu $\bar{x}_n$ eta x_{n+1} erabiliz nola kalkula daitekeen $\bar{x}_{n+1}$.
    b) Frogatu ezazu: $(n+1)s^2_{n+1} = ns^2_n + \frac{n}{n+1}(\bar{x}_n - x_{n+1})^2$.

16. Zerbitzariak eskari batetik bestera pasatutako denboren (s) 55 behaketa dituzu jarraian:

1.86	3.93	4.19	3.35	3.32	1.96	4.51	4.98	4.11	3.24
2.92	3.75	3.49	3.24	3.01	1.90	4.89	3.00	3.46	4.48
3.33	4.08	3.99	3.03	3.18	1.88	3.99	3.05	3.04	2.87
4.68	4.41	3.06	3.67	3.45	2.43	3.34	2.99	2.92	4.19
4.28	3.45	4.12	3.37	3.53	2.36	3.15	4.47	5.85	2.91
3.57	3.20	3.54	3.61	3.29					

    Laburtu itzazu datuok atal honetan erabilitako teknikak erabiliz, eta azaldu itzazu idatziz zure laburpenaren ezaugarri nagusiak.

17. Honako irudi hauek aldagaiaren balioak 4 laginetan nola banatu diren erakusten dute. Banaketaren forman erreparatuz eta hori erreferentziatzat hartuta, bururatzen al zaizu irudi horiek ordenatzeko modurik? Arrazoitu zure erantzuna.



18. Azken urteetako hirien kezka nagusietako bat da zabor-bilketa eta zaborren deuseztea. Hori dela eta, Euskal Autonomi Elkarteko 20 hiritan azken hilabetean bildutako zabor kantitatea (Tn) aztertu nahi da. Hau da bildutako informazioa:

21	22	28	30	32	35	39	47	48	49
55	56	60	62	63	66	73	76	89	215

   a) Kalkula itzazu ezagutzen dituzun joera zentraleko estatistiko guztiak, eta aukeratu datu horiek ordezkatzeko egokiena. Arrazoitu zure erantzuna.
   b) Demagun datuak tarteetan biltzea erabaki dela honako era honetara: 20 zabalerako hiru tarte, 20 tonatan hasita, eta azken tarte bat (zuk nahi bezala). Horren baitan, kalkula itzazu sakabanapen absolutua eta erlatiboa neurtzeko estatistiko bana, estatistikoen izena adieraziz.
   c) Zein hiri-portzentaje dago [50, 55) tartean? Kalkuluak egin itzazu datuak elkartuta eta elkartu gabe. Komentatu zure erantzuna.

19. Datuak eskuratzea gero eta errazagoa da. Hor dago adibidez *Wikidata*[2]. Hortik hartuta, Wikipedian erregistratuta dagoen lurraldeen populazioa duzu hemen. Irudika itzazu grafiko batean, eta kalkulatu mediana eta 10. eta 90. pertzentilak. Erabili eskura duzun softwarea.

[2] https://eu.wikipedia.org/wiki/Wikidata

```r
# Datuak deskargatu ditugu query.csv izeneko fitxategi batean,
# Wikidatatik
datuak <- read.table("query.csv", header=TRUE, sep=",")
populazioa <- datuak$population
# populazioa objetuan dituzu datuak.
```

4. gaia

ESTATISTIKA DESKRIBATZAILEA. Bi dimentsiotako banaketak. Korrelazioa eta erregresioa.

1. Ikasle talde batek *Datu-egiturak eta algoritmoak (DEA)* eta *Zenbakizko konputazioa (ZK)* irakasgaietan lortutako kalifikazioak honako hauek dira:

DEA	3	4	6	7	5	8	7	3	5	4	8	5	5	8	8	8	5
ZK	5	5	8	7	7	9	10	4	7	4	10	5	7	9	10	5	7

   a) Zenbat ikasleren datuak bildu dira?
   b) Eratu ezazu maiztasu-taula.
   c) Azaldu ezazu informazio hori sakabanapen-grafikoan. Horren arabera, nolako korrelazioa dagoela esango zenuke?
   d) Kalkula itzazu bazter-banaketen batezbesteko eta bariantzak.
   e) Kalkula ezazu kobariantza.
   f) Kalkula ezazu korrelazioa.

R erabiliz garatzeko:

```r
# Sartu datuak
x <- c(3, 4, 6, 7, 5, 8, 7,  3, 5,  4,  8,  5, 5,  8,  8, 8, 5)
y <- c(5, 5, 8, 7, 7, 9,  10, 4, 7, 4, 10, 5, 7, 9, 10, 5, 7)
n <- length(x)

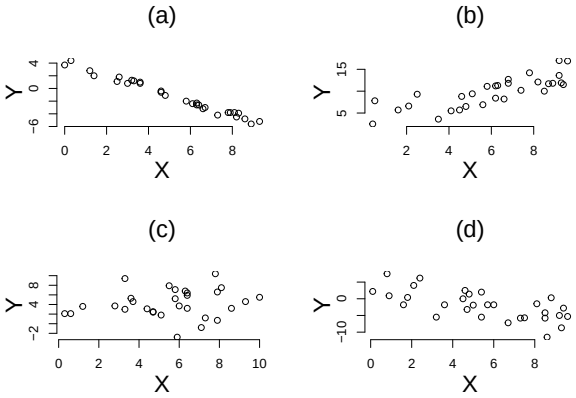
# x (DEA)ren batezbestekoa eta bariantzarako baturak
sum(x)
sum(x^2)
SXX <- sum(x^2) - (sum(x)^2)/n
# y (ZK)rentzako
sum(y)
sum(y^2)
SYY <- sum(y^2) - (sum(y)^2)/n

# Kobariantzarako
# zer dago x*y bektorean?
x*y
```

```r
sum(x*y)
# Berdinak dira sum(x*y) eta sum(x)*sum(y)?
#Idatzi bakoitzari dagokion espresio matematikoa
SXY <- sum(x*y) - sum(x)*sum(y)/n
SXY/(n-1)

# Korrelazioa
SXY/sqrt(SXX*SYY)
# edo zuzenean
cor(x, y)
```

2. Lotu korrelazio-zenbakiak irudiekin: $r_1 = -0.990$, $r_2 = -0.667$, $r_3 = 0.157$ eta $r_4 = 0.802$.



3. Irudika ezazu adibide bat, non $r > 0$ baita, baina puntu bat kentzean $r \approx 0$ bihurtzen baita.

4. Kalkulu-zentro batean uste dute badagoela zerikusirik prozesatu beharreko lan kopuruaren (X) eta sistemaren «erortze'» kopuruaren (Y) artean. Uste hori egiaztatu edota gezurtatzeko datuak bildu zituzten 25 hilabeteren buruan:

X	Y	X	Y	X	Y	X	Y	X	Y
1749	3	1601	4	1445	2	1355	1	1703	5
1624	6	2001	4	1237	3	1344	4	1648	3
1623	3	1598	4	1788	3	1667	3	1546	2
1456	2	1488	1	1729	2	1656	3	1621	1
1492	0	1522	5	1769	6	1332	2	1678	4

a) Adieraz itzazu datuak sakabanapen-grafikoan. Horren arabera, kalkuluzentroaren susmoak oinarririk baduela uste duzu?
b) Kalkula ezazu bi aldagaien arteko kobariantza.
c) Kalkula ezazu korrelazioa.
d) Kalkula ezazu X-rekiko Y-ren erregresio zuzenaren ekuazioa, eta marraztu sakabanapen-grafikoan.
e) Hilabete batean 1 800 prozesatu behar badira, batez beste zenbat alditan «erortzea» espero da eredu lineal horren arabera?
f) Eredu horren arabera, Y-ren aldakortasunaren zein proportzio azaldu daiteke X-ri esker?

5. Irudi bat memorian kargatzeko behar den denbora (s) (D) eta irudiaren tamaina (MB) (T) batera aztertzeko datuak bildu ditugu.

T:	11	4	4	9	7	5	16	4	13	10	12	8	15	10	11
D:	1.4	0.1	0.1	1.2	0.5	0.6	2.0	0.3	1.6	1.5	1.7	1.1	1.6	1.3	0.8

a) Eraiki ezazu grafiko egokia bi aldagaien artean erlaziorik badagoen jakiteko. Grafikoa ikusita, uste duzu erlaziorik badagoela bi aldagaien artean? Arrazoitu zure erantzuna.
b) Kalkula ezazu *irudiaren tamaina*ren gaineko *irudia memorian kargatzeko behar den denbora*ren erregresio zuzena. Eredu horren arabera, zenbat denbora espero behar dugu 10MB tamainako irudi bat memoriara kargatzeko?
c) Eredu horren arabera, irudia kargatzeko behar den denboraren aldakortasunaren zein proportzio azaldu daiteke irudiaren tamaina ezagutzen badugu?

5. gaia

PROBABILITATE KALKULUA. Oinarrizko probabilitate-teoria.

1. Txanpon bat hiru aldiz bota da. Kalkula itzazu:
   - Lagin-espazioa.
   - Adieraz itzazu honako gertaera hauek oinarrizko gertaeren bidez:
     - «Hiru jaurtiketek emaitza bera eman dute».
     - «Emaitza berdina bi aldiz soilik agertu da».
     - «Gutxienez bi aurpegi agertu dira».
     - «Zehazki bi aurpegi agertu dira».
     - «Lehen bi jaurtiketetan aurpegia atera da».

2. Sei irudi-txartel ditugu, eta badakigu hauetako bi akastunak direla. Txartelak banan-banan probatuko ditugu akastun bat aurkitu arte. Deskriba ezazu esperimentuari dagokion lagin-espazioa.

3. 1etik 10era zenbaki bat zoriz aukeratuko dugu. Demagun $A \equiv$ *«Aukeratutako zenbakia bakoitia da»*, $B \equiv$ *«Aukeratutako zenbakia lehena da»* eta $C \equiv$ *«Aukeratutako zenbakia 3ren multiploa da»* gertaerak kontuan hartzen ditugula. Adieraz itzazu honako gertaera hauek lagin-espazioaren azpimultzo gisa:

$$A \cap B \quad A - B \quad A \cup B \quad \bar{A} \cap \bar{B} \quad A \cap C \quad A - C \quad B \cap C \quad \bar{A} \cap C \quad (A \cup B) \cap \bar{C}$$

4. Demagun A, B eta C hautazko gertaerak direla. Aurki itzazu honako gertaera hauentzat A, B eta C-ren menpeko adierazpenak:

   a. A soilik gertatzen da.
   b. A eta B gertatzen dira, baina ez C.
   c. Hiru gertaerak gertatzen dira.
   d. Gutxienez gertaera bat gertatzen da.
   e. Gutxienez bi gertaera gertatzen dira.
   f. Bat, eta soilik bat gertatzen da.
   g. Bi, eta soilik bi gertatzen dira.
   h. Bat ere ez da gertatzen.
   i. Bi baino gehiago ez dira gertatzen.

5. Jo dezagun bi gertaeren probabilitateak $P(A) = 1/3$ eta $P(B) = 1/2$ direla. Kalkulatu $\bar{A} \cap B$ gertaerari dagokion probabilitatea baldintza desberdinetan:

   a. A eta B bateraezinak direnean.
   b. A B-ren parte denean.
   c. $P(A \cap B) = 1/8$ denean.

6. Ordenagailuen inguruko aholkularitza-enpresa batek hiru proiektutan esku hartu du. Izan bitez $A_i \equiv$ *«i. proiektua eskuratu du»*, $i = 1, 2, 3$ eta $P(A_1) = 0.22$, $P(A_2) = 0.25$, $P(A_3) = 0.28$, $P(A_1 \cap A_2) = 0.11$, $P(A_1 \cap A_3) = 0.05$, $P(A_2 \cap A_3) = 0.07$, $P(A_1 \cap A_2 \cap A_3) = 0.01$. Honako gertaera hauen esanahia hitzez adieraz ezazu, eta kalkulatu bakoitzari dagokion probabilitatea.

   a. $A_1 \cup A_2$
   b. $\bar{A}_1 \cap \bar{A}_2$
   c. $A_1 \cup A_2 \cup A_3$
   d. $\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3$
   e. $\bar{A}_1 \cap \bar{A}_2 \cap A_3$
   f. $(\bar{A}_1 \cap \bar{A}_2) \cup A_3$

7. Hiru produkzio-katedun fabrika batek A, B eta C produktu desberdinak ekoizten ditu. Produktu bakoitza produkzio-kate batek ekoitziko du.

   a. Eraiki ezazu 3 produkzio-kateetan produktuen esleipenerako dauden 6 era desberdinak adierazten dituen zuhaitz-diagrama.
   b. Esleipena zorizkoa izan bada, zuhaitz-diagramaren bide bakoitza ekiprobablea izango da. Zein da lehen produkzio-katean A produktua erabilia izateko probabilitatea?

8. Inkesta baten arabera, graduko ikasleen % 50ek A irakasgaia aukeratzen du, % 65ek B irakasgaia eta % 85ek bietako bat gutxienez.

    a. Zenbatek aukeratzen dituzte biak?
    b. Zenbatek bat bera ere ez?

9. Kutxa batean 50 bola daude. Badira gorriak eta handiak, urdinak eta handiak, eta urdinak eta txikiak. 30 handiak dira, eta 25 urdinak. Zenbat dira urdinak eta handiak?

10. Dado bat kargatuta dugu, zenbaki bakoiti bat lortzeko aukerak bikoiti bat lortzekoaren bikoitza izateko eran. Dadoa behin botako dugu.

    a. Kalkula ezazu 3 baino zenbaki handiagoa lortzeko probabilitatea.
    b. Kalkula ezazu lortutako emaitza karratu perfektua izateko probabilitatea.
    c. Kalkula ezazu lortutako emaitza karratu perfektua izateko probabilitea, jakinik 3 baino emaitza handiagoa lortu dugula.

11. Elektrogailu-denda batean 3 telebista-marka saltzen dituzte. Salmenten % 50 M1 markakoa (merkeena) izaten da, % 30 M2 markakoa eta % 20 M3 markakoa. M1 markako telebisten % 25ak konponketaren bat behar du bere garantia-aldian; era berean, M2 eta M3 markako telebisten % 20ek eta % 10ek, hurrenez hurren, behar dute konponketaren bat.

    a. Zein da erosleak M1 markako eta konponketa beharko duen telebista erosteko probabilitatea?
    b. Zein da erosleak konponketa beharko duen telebista bat erosteko probabilitatea?
    c. Demagun eroslea garantia–aldian konponketa behar duen telebistarekin itzuli dela dendara. Zein da M1 markako telebista izateko probabilitatea? Eta M2 markakoa? Eta M3 markakoa?

12. K1 kutxan 6 pelota gorri eta 4 berde daude, eta K2 kutxan berriz 4 pelota gorri eta 3 berde. K1 kutxatik zoriz pelota bat aukeratzen da, eta K2 kutxan sartu. Ondoren, K2 kutxatik beste pelota bat zoriz aukeratzen da, eta atzera K1 kutxan sartu.

    a. Zein da K1 kutxan pelota gorria eta K2 kutxan pelota gorria aukeratzeko probabilitatea?
    b. Prozesua amaitu ondoren, zein da bi kutxetan hasierako pelota-proportzioa mantentzeko probabilitatea?

13. A eta B gertaerak emanik ($P(B) > 0$), froga ezazu: $P(A|B) + P(\bar{A}|B) = 1$.

14. A, B eta C gertaerak emanik ($P(C) > 0$), froga ezazu: $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$.

15. A eta B gertaerak emanik ($P(B) > 0$), froga ezazu $P(A) = P(B)P(A|B) + P(\bar{B})P(A|\bar{B})$.

16. Gasolindegi jakin batean, bezeroen % 40ek 95 oktanotako berunik gabeko gasolina (BG95) erabiltzen dute, % 35ek gasoila (G) eta % 25ek 98 oktanotako berunik gabeko gasolina (BG98). BG95 erabiltzen dutenen artean % 30ek depositua goraino betetzen du, G erabiltzen dutenen artean % 60ek eta BG98 erabiltzen dutenen artean, berriz, % 50ek.

    a. Zein da bezeroak G gasoila eskatu eta depositua betetzeko probabilitatea?
    b. Zein da bezeroak depositua betetzeko probabilitatea?
    c. Baldin eta bezeroak depositua bete badu, zein da BG95 eskatua izateko probabilitatea? Eta G? Eta BG98?

17. Aurreko ariketaren haritik, demagun:

    - BG95 eta depositua betetzen dutenen artean, % 70ek kreditu-txartelarekin ordaintzen du.
    - BG95 erabili baina depositua betetzen ez dutenen artean, % 50ek kreditu-txartelarekin ordaintzen du.
    - G eta depositua betetzen dutenen artean, % 60ek kreditu-txartelarekin ordaintzen du.
    - G erabili baina depositua betetzen ez dutenen artean, % 50ek kreditu-txartelarekin ordaintzen du.
    - BG98 eta depositua betetzen dutenen artean, % 50ek kreditu-txartelarekin ordaintzen du.
    - BG98 erabili baina depositua betetzen ez dutenen artean, % 40ek kreditu-txartelarekin ordaintzen du.

    Kalkula itzazu honako gertaera hauen probabilitateak:

    a. G, depositua bete eta kreditu-txartela erabili.
    b. BG98, depositua ez bete eta kreditu-txartela erabili.
    c. BG98 eta kreditu-txartela erabili.
    d. Depositua bete eta kreditu-txartela erabili.
    e. Kreditu-txartela erabili.
    f. Baldin eta bezeroak kreditu-txartela erabili badu, zein da BG98 eskatua izateko probabilitatea?

18. Baldin eta uretan B motako bakterioa badago, testak positiboa emateko probabilitatea 0.9 da. Bakteriorik ez badago, testak negatiboa emateko probabilitatea 0.8 da. Ur-laginak benetan B motako bakterioak edukitzeko probabilitatea 0.2 dela jakinik, kalkula ezazu:

    a. Testa positiboa ateratzeko probabilitatea.
    b. Testa positiboa izan denean, ur-laginak B motako bakterioak edukitzeko probabilitatea.

19. Lurralde jakin batean desagertzen diren hegazkin arinen % 70 aurrerago aurkitu egiten dira. Aurkitutako hegazkinen % 60ak larrialdi-lokalizatzailea du;

aldiz, aurkitu ez direnen % 90ek ez. Demagun hegazkin arin bat desagertu dela:

a. Baldin eta larrialdi-lokalizatzailea badu, zein da hegazkina ez aurkitzeko probabilitatea?
b. Baldin eta larrialdi-lokalizatzailerik ez badu, zein da hegazkina aurkitzeko probabilitatea?

20. Etxeko nagusiak atea irekitzeko hiru giltza dauzka. Giltzak guztiz bereizgaitzak dira, eta beren arteko bakar batek irekitzen du bere etxeko atea. Behar duen giltza aurkitzeko, honako metodo hauetakoren batez baliatzen da:

I) Giltzak bata bestearen ondoren erabiltzen ditu, giltza berbera bitan ez erabiltzeko kontua edukiz.
II) Giltza bat erabiltzen du eta, ez badu atea irekitzen, giltzatakoa astindu eta berriro saiatzen da. Horrela, giltza bera bi aldiz erabiltzeko arriskua du.

Honako hau eskatzen da:

a. I metodoa jarraituz, zein da atea hirugarren saiakeran irekitzeko probabilitatea? Eta II metodoa jarraituz?
b. Etxeko nagusia lasai dabilenean (hori hiru egunetik behin gertatzen da) II metodoa erabiltzen du, eta presaka dabilenean I metodoa. Bi saiakeretan huts egin duela jakinez gero, zein da nagusia lasai ibiltzeko probabilitatea?

21. A motako osagaiak ekoizten dituen enpresa batek osagaiak 10naka osaturiko sortetan bidaltzen dizkio bezeroari. Demagun sorta horietako % 50ek ez duela akastun osagairik, % 30ek akastun bakarra du, eta % 20k, berriz, bi akastun. Erosleak sorta bateko 2 osagai zoriz aukeratu eta frogatu egiten ditu. Kalkula ezazu sortan 0, 1 eta 2 akastun egoteko probabilitatea, baldin eta:

a. Frogatutako osagaiek ez bazuten akatsik.
b. Frogatutako osagai bat akastuna bazen.

22. Zentsuko datuetatik abiatuz, begi belzdun aitek eta begi belzdun umeek zentsuan dauden pertsonen % 5 osotzen dute; begi belzdun aitek eta begi argidun umeek % 7.9; begi argidun aitek eta begi belzdun umeek % 8.9; eta begi argidun aitek eta umeek % 78.2. Azter ezazu aitaren begien koloreak nola baldintzatzen duen semearen begien kolorea.

23. Hiri handi batean bizi den lagun batek bi auto ditu, bata txikia eta bestea handia. Lanera joateko, lautik hirutan auto txikia erabiltzen du, eta orduan aparkatzeko arazorik izaten ez duenez, lanera garaiz iristen da 0.9ko probabilitatearekin. Auto handiarekin joaten denean, ordea, aparkatzeko arazoak direla eta, lanera garaiz iristeko probabilitatea 0.6 da. Lanera garaiz iritsi dela jakinik, zein da auto txikiarekin joana izateko probabilitatea?

24. Suposa dezagun zoriz pertsona bat aukeratzen dela, eta kontsidera ditzagun honako gertaera hauek: $A \equiv$ *«Aukeratutako pertsona 2m luze baino altuagoa da»*, $B \equiv$ *«Aukeratutako pertsona saskibaloiko jokalari profesionala da»*.

    Zein izango da altuagoa, $P(A|B)$ ala $P(B|A)$? Zergatik?

25. Tuberkulosiaren presentzia detektatzeko proba-diagnostiko bat aplikatu da ehungintzako langile talde batean. Orain arte egindako ikerketen arabera, mota honetako langileetan gaixotasunaren prebalentzia (gaixorik dauden pertsonen proportzioa) % 14koa dela onartzen da. Ikerketa berdinetan, langile osasuntsuen % 17k positibo eta langile gaixoen %5ek negatibo eman duela behatu da. Kalkula ezazu osasuntsu dauden eta positibo eman duten langileen proportzioa; eta gaixorik dauden eta negatibo eman duten langileen proportzioa.

26. Froga ezazu A eta B gertaerak askeak badira, orduan $\bar{A}$ eta B ere askeak direla.

27. Probabilitate-espazioan A eta B gertaerak askeak badira, froga ezazu orduan $\bar{A}$ eta $\bar{B}$ ere askeak direla.

28. Izan bitez A eta B gertaerak, non $P(A) = 0.4$ eta $P(A \cup B) = 0.7$ diren. Zenbat balio behar du $P(B)$-k, A eta B zera izan daitezen:

    - Bateraezinak.
    - Independenteak.

29. A eta B gertaerak bateraezinak dira, eta probabilitate positiboa dute. Independenteak al dira?

30. A eta B gertaerak bateraezinak dira. Zein baldintzapetan dira bateraezinak $\bar{A}$ eta $\bar{B}$?

    A eta B gertaerak askeak dira. Zein baldintzapetan dira askeak $\bar{A}$ eta $\bar{B}$?

31. Demagun kalkulu-zentro bateko zuzendariak bilerak goiz eta arratsalde izaten dituela. Izan bitez: $A \equiv$ *«Goizeko bilerara garaiz iritsi da»*, $B \equiv$ *«Arratsaldeko bilerara garaiz iritsi da»*.

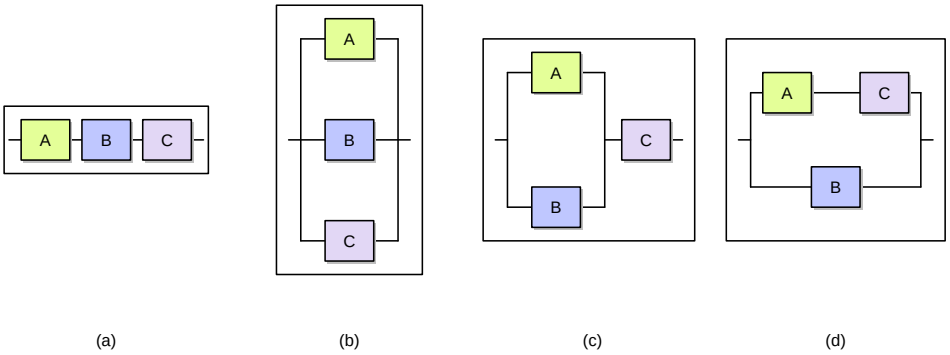
    a. Baldin eta $P(A) = 0.4$, $P(B) = 0.5$ eta $P(A \cap B) = 0.25$ badira, A eta B gertaerak askeak al dira?
    b. Demagun A eta B gertaerak askeak direla, $P(A) = 0.4$, $P(B) = 0.5$ izanik. Zein da zuzendaria bi bileretara garaiz iristeko probabilitatea? Eta zehazki bakar batera garaiz iristekoa?

32. Bi kaxa ditugu, eta bakoitzak bi kaxoi ditu. A kaxak urrezko txanpon bat du kaxoi batean, eta zilarrezkoa bestean. B-k, aldiz, urrezkoa bietan. Zoriz kaxa bat eta zoriz kaxoi bat aukeratu eta gero, urrezko txanpona suertatu da. Kalkulatu B kaxatik hartua izateko probabilitatea.

33. Enpresa jakin batek gauez bidaltzen dituen paketeen % 40 $E1$ izeneko mezularien enpresaren bitartez egiten du, eta pakete horien % 2a soilik iristen da berandu. Gauez bidaltzen dituen paketeen % 50, berriz, $E2$ mezularien bitartez eta % 10 $E3$ mezularien bitartez egiten du. Azken horiekin bidalitakoen % 1 eta % 5 iristen dira berandu, hurrenez hurren. Kalkula itzazu:

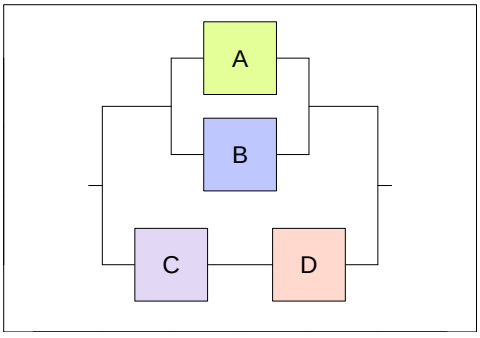
    a. Zoriz aukeratutako pakete bat $E1$ mezularien bidez bidali eta berandu iristeko probabilitatea.
    b. Zoriz aukeratutako pakete bat berandu iristeko probabilitatea.
    c. Zoriz aukeratutako pakete bat garaiz iritsi dela jakinik, kalkula ezazu $E1$ mezularien bitartez bidalia ez izateko probabilitatea.

34. Hiru osagaiz osaturiko hainbat sistema dugu. Sistemak azpiko 10.1 irudian agertzen diren bezala antolaturik daude. Osagai guztien funtzionamendua askea dela jakinik eta bakoitzak ongi funtzionatzeko probabilitatea 0.95 izanik, kalkula ezazu sistema bakoitzaren fidagarritasuna (*fidagarritasuna*: sistema osoak ongi funtzionatzeko probabilitatea).

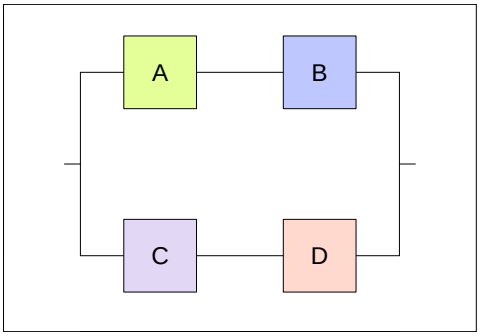


10.1 irudia. 3 osagaiekin osatutako 4 sistema.

35. Azpiko irudian erakusten den bezala, 4 osagaidun sistema bat dugu. Osagai bakoitzaren funtzionamendua besteenarekiko askea da, eta bakoitzak ongi funtzionatzeko probabilitatea 0.9 da. Zein da sistemaren fidagarritasuna?



36. 4 osagaidun sistema bat dugu azpiko irudian agertzen den bezala. Osagai bakoitzaren funtzionamendua besteenarekiko askea da, eta sistemaren fidagarritasuna 0.99 izatea nahi bada, zein da osagai bakoitzak ongi funtzionatzeko bermatu beharko dugun probabilitatea?



37. Enpresa banatzaile jakin batera iristen diren osagaiak bi ikuskarik aztertzen dituzte (osagai bakoitza bi ikuskarik aztertzen dute). Ikuskari bakoitzak akastun osagai guztien % 90a soilik identifikatzen du, eta gutxienez ikuskari batek ez du identifikatzen akastun osagai guztien % 20. Kalkulatu honako probabilitate hauek:
    (a) Akastun osagai bat lehen ikuskariak bakarrik identifikatzea.
    (b) Akastun osagai bat bietako ikuskari bakarrak identifikatzea.
    (c) Hiru osagai akastun ez ikuskari batek eta ez besteak ez identifikatzea.

38. Autoen aseguru-etxe batek gidariak hiru multzotan sailkatzen ditu: arrisku txikia (A1), arrisku ertaina (A2), arrisku handia (A3). Aseguratuen artean % 30 arrisku txikikoak, % 50 arrisku ertainekoak eta % 20 arrisku handikoak dira. Urtebetean gidariak istripu bat izateko probabilitatea 0.1, 0.3 edo 0.5 da, baldin eta arrisku txikikoa, ertainekoa edo handikoa bada, hurrenez hurren. Gidari batek istripu bat izan badu, zein da gidaria arrisku txikikoa izateko probabilitatea? Eta arrisku ertainekoa?

39. Zorro batean hiru txanpon daude. Bat gorria da bi aldeetatik (T_1), beste bat urdina bi aldeetatik (T_2), eta hirugarrena gorria alde batetik eta urdina bestetik (T_3). Txanpon bat zoriz aukeratua izan da, eta ezkutuan n aldiz botaz gero guztietan gorria dela ikusi da. Zein da beste aldea ere gorria izateko probabilitatea?

6. eta 7. gaiak

PROBABILITATE KALKULUA. Zorizko aldagaiak eta probabilitate-banaketak. Laginketa, simulazioa eta limitearen teoremak.

Aldagai diskretuak

1. Eman itzazu Bernoulliren banaketa jarraituko duten hiru zorizko aldagairen adibideak.

2. Izan bedi $X \equiv$ *«Posta-kodean, 0ren desberdinak diren digitu kopurua»*. Zein balio hartu ditzake aldagai honek? Eman itzazu 3 posta-kode eta bakoitzari dagokion aldagaiaren balioa.
   Oharra: posta-kodeak 5 digitu ditu.

3. Ondoren definitutako zorizko aldagai (z.a.) bakoitzaren heina zein den esan ezazu, baita aldagaia diskretua den ala ez ere.

   a. $X \equiv$ *«Dozenako kutxan hautsirik ez dagoen arrautza-kopurua»*
   b. $Y \equiv$ *«Golfean hasi berri denak pilota jotzeko egin behar duen saiakera-kopurua»*
   c. $U \equiv$ *«Parke natural batetako sugegorri baten luzera»*
   d. $T \equiv$ *«Fakultateko idazkaritzako leihatilan itxaron beharreko denbora»*
   e. $Z \equiv$ *«Txanpon bakar bat airera bota beharreko aldi-kopurua hiru aldiz jarraian emaitza berdina lortu arte»*

4. Kalkulatu honako funtzio hauetan a-ren balioa, ahal bada, probabilitate-balioak izan daitezen. Ondoren, irudika itzazu barra-diagrama bidez.

   a. $P(X = k) = ak\,, \quad k = 1, 2, 3, 4, 5.$
   b. $P(X = k) = ak\,, \quad k = 1, 2, \ldots, n.$
   c. $P(X = k) = a/k\,, \quad k = 1, 2, \ldots.$
   d. $P(X = k) = a/k^2\,, \quad k = 1, 2, \ldots.$

5. Izan bedi $X \equiv$ *«Auto batean presio baxuegia duen gurpil kopurua»*.

   Honako hiru funtzio hauetan, baten bat Xren probabilitate-legea (p.l.) da. Zein, ordea? Zergatik? Irudika ezazu barra-diagrama bidez.

x	0	1	2	3	4
$f(x)$	0.3	0.2	0.1	0.05	0.05
$g(x)$	0.4	0.1	0.1	0.1	0.3
$h(x)$	0.4	0.1	0.2	0.1	0.3

   a. Aurreko atalean oinarrituz, kalkula itzazu: $P(2 \leq X \leq 4)$, $P(X \leq 2)$ eta $P(X \neq 0)$.
   b. Auto batean zein da batez beste espero izatekoa den presio baxuegiko gurpil kopurua?

6. Egiaztatu honako hauek probabilitate-legeak direla, eta kalkulatu bakoitzaren espero izateko balioa eta bariantza.

   a. $P(X=-2)=0.4$, $P(X=0)=0.1$, $P(X=1)=0.3$, $P(X=4)=0.2$.
   b. $P(X=k)=1/6\cdot k\,, \quad k=1,2,3.$
   c. $P(X=k)=\binom{10}{k}\cdot\frac{1}{2^{10}}\,, \quad k=0,1,2,\ldots,10.$

7. Demagun bi dado botako ditugula. Izan bedi $M\equiv$ *«bi dadoetako emaitzen maximoa»* (adibidez, $M(2,5)=5$).

   a. Zein da M-ren probabilitate-legea? (*Iradokizuna:* lehengo kalkulatu $f(1)$, $f(2)$,...)
   b. Kalkulatu eta irudikatu M-ren banaketa-funtzioa.

8. Aseguru-etxe batek era desberdinetara ordaintzeko aukera ematen die bere bezeroei. Izan bedi $X\equiv$ *«bi ordainketa jarraituen arteko hilabete kopurua»* eta bere banaketa funtzioa honako hau:

$$F(x)=\begin{cases} 0\,, & x<1 \\ 0.30\,, & 1\le x<3 \\ 0.40\,, & 3\le x<4 \\ 0.45\,, & 4\le x<6 \\ 0.60\,, & 6\le x<12 \\ 1\,, & x\ge 12 \end{cases}$$

   a. Zein da X-ren probabilitate-legea?
   b. Banaketa-funtzioa soilik erabiliz, kalkula itzazu $P(3\le X\le 6)$ eta $P(X\ge 4)$.

9. X zorizko aldagaiak honako era honetara definitutako probabilitate-legea du:

$$f(x)=\begin{cases} k & \text{baldin eta } x=0, \\ 2k & \text{baldin eta } x=1, \\ 3k & \text{baldin eta } x=2. \end{cases}$$

   Kalkula itzazu:

   a. k-ren balioa benetan p.l. izateko.
   b. $P(X\le 2)$ eta $P(0<X<2)$.
   c. X-ren banaketa-funtzioa, itxaropen matematikoa eta bariantza.

10. Lan jakin bat egiteko n hautagai ditugu, eta 1, 2, ..., n moduan izendatu dira. Izan bedi $X\equiv$ *«lana egingo duen hautagaiari izendatutako zenbakia»* eta hautagaia zoriz aukeratuko denez:

$$f(x)=\begin{cases} 1/n\,, & x=1,2,\ldots,n \\ 0\ , & \text{bestela} \end{cases}$$

    Banaketa honi banaketa uniforme diskretua esaten zaio, $X\sim\mathcal{UD}(n)$. Kalkula itzazu banaketa honen itxaropen matematikoa eta bariantza.

Oharra. Kontuan hartu: $\sum_{i=1}^{n} i = \dfrac{n(n+1)}{2}$, $\sum_{i=1}^{n} i^2 = \dfrac{n(n+1)(2n+1)}{6}$.

11. Kiosko bateko nagusiak aldizkari jakin bateko aleak astero eskatu behar ditu. Izan bedi $X \equiv$ *«astean zehar aldizkaria erostera etorritako bezero kopurua»* eta honakoa bere p.l.:

$$f(x) = \begin{cases} 1/15\,, & x = 1 \\ 2/15\,, & x = 2 \\ 1/5\,, & x = 3 \\ 4/15\,, & x = 4 \\ 1/5\,, & x = 5 \\ 2/15\,, & x = 6 \\ 0\,, & \text{bestela} \end{cases}$$

    a. Demagun nagusiak 0.25 euro ordaintzen duela eskatutako aldizkariaren ale bakoitzeko, baina saldu, euro 1ean saltzen duela. Zer komeni zaio gehien nagusiari, astero aldizkariaren 3 ale ala 4 ale erostea?

    Iradokizuna: adieraz itzazu nagusiak izango dituen irabaziak kasu bakoitzerako.

    b. Irudika ezazu X-ren probabilitate-legeari dagokion barra-diagrama. Kalkula ezazu $-X$-ren probabilitate-legea eta irudikatu dagokion barra-diagrama. Bi irudi horiek alderatuz, zer esan daiteke $VAR(X)$ eta $VAR(-X)$-ri buruz?

12. Demagun $X \equiv$ *«hiri batean familiako haur kopurua»* neurtzen duen z.a.-k honako probabilitate-lege honi jarraitzen diola:

$$f(x) = \begin{cases} 0.47\,, & x = 0 \\ 0.30\,, & x = 1 \\ 0.10\,, & x = 2 \\ 0.06\,, & x = 3 \\ 0.04\,, & x = 4 \\ 0.02\,, & x = 5 \\ 0.01\,, & x = 6 \\ 0\,, & \text{bestela} \end{cases}$$

    a. Kalkula ezazu X-ren banaketa-funtzioa.
    b. Kalkula ezazu X-ren itxaropen matematikoa. Zein da zenbaki horren esanahia?
    c. Kalkula itzazu X-ren bariantza eta desbideratze estandarra.
    d. Suposa dezagun hiri honetako udalak haur bakoitzeko 12 euro ordaintzen dizkiola familiari. $Y = 12X$ z.a. definitzen badugu, zer adierazten du Y-k? Zein da Y-ren probabilitate-legea?
    e. Kalkula itzazu Y-ren itxaropena, bariantza eta desbideratze estandarra.

13. Lurralde jakin bateko nekazariak badaki aurretiko esperientzian oinarrituz, udako eguraldia ohikoa bada, 80 000 euroko irabazia izango duela. Badaki,

baita ere, uda bereziki lehorra joaten bada irabaziak 50 000 euro izango direla; baina irabaziak 20 000 euro izango direla ekaitz gogorren batek uzta hondatzen badu. Meteorologia-agentziaren arabera, ohiko uda izateko probabilitatea 0.70 da, uda bereziki lehorra izatekoa 0.20, eta ekaitz gogorrak pairatzekoa 0.10. Izan bedi X datorren udan nekazariak izango dituen irabaziak adierazten dituen z.a.

a. Jar ezazu taula batean X-ren probabilitate-legea.
b. Zein da irabaziak 50 000 euro edo txikiagoak izateko probabilitatea?
c. Zein da espero izateko batez besteko irabazia?
d. Demagun nekazariak 3 000 euroko aseguru bat kontrata dezakeela. Horren arabera, baldin eta ekaitz gogorren batek uzta hondatzen badu, aseguru-etxeak 20 000 euro ordainduko lizkioke, baina gainontzean ez lioke ezer ordainduko. Aurkitu ezazu aseguru hori kontratatuz gero nekazariak izango dituen irabazien probabilitate-legea. Zein da espero izateko batez besteko irabazia? Zure ustez, interesatzen al zaio aseguru hori kontratatzea?

14. X z.a. emanik, honako desberdintza hau betetzen da (Tchebychev-en desberdintza)

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}\,, \qquad \text{non } EX=\mu \text{ eta } VAR(X)=\sigma^2 \text{ baitira.}$$

Hau da, X aldagaia bere batezbestekotik k desbideratzeetara baino urrunago egoteko probabilitatea $1/k^2$ baino txikiagoa da.

a. Zein da goi-borne horren balioa $k=2$ denean? Eta $k=3$? $k=4$? $k=5$?
b. Ordenadore-enpresa batek 6 telefono-linea ditu, eta X-k momentu jakin batean zenbat linea batera erabiltzen ari den adierazten du. Hau da dagokion probabilitate-legea:

x	0	1	2	3	4	5	6
$f(x)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

Kalkula itzazu kasu honetarako $P(|X-\mu| \geq k\sigma)$ probabilitateak a. ataleko k-ren balioetarako. Aipaturiko desberdintzaren zehaztasuna nolakoa dela iruditzen zaizu?

c. Demagun X-k hiru balio bakarrik har ditzakeela, -1, 0 eta 1; 1/18, 8/9 eta 1/18 probabilitateekin, hurrenez, hurren. Kalkula ezazu $P(|X-\mu| \geq 3\sigma)$ eta konpara ezazu dagokion goi-bornearekin.

15. Joko batean hiru dira litezkeen emaitzak: 100 diru-unitate irabazi 1/5 probabilitateaz, 40 galdu 2/5 probabilitateaz, eta 0 irabazi edo galdu bestela. Kalkulatu espero izateko balioa.

16. $X \sim Bin(n,p)$ izanik, froga ezazu probabilitate-legea izateko $\sum_x P(X = x) = 1$ baldintza betetzen duela.

    Laguntza: $(a+b)^k = \sum_{i=0}^{k} \binom{k}{i} a^i b^{k-i}$, $\quad a, b \in \mathbb{R}$, $\quad k \in \mathbb{N}$.

17. Definizioa aplikatuz edo probabilitate-taulak erabiliz, kalkula itzazu honako probabilitate hauek:

    a. $X \sim Bin(8, 0.3)$ izanik: $P(2 \leq X \leq 4)$, $P(X \geq 2)$, $P(2 < X < 6)$.
    b. $X \sim Bin(8, 0.7)$ izanik: $P(X \leq 1)$.
    c. $X \sim Geo(0.2)$ izanik: $P(X = 4)$, $P(X \leq 5)$, $P(X > 3)$.
    d. $X \sim H(15, 11, 5)$ izanik: $P(X = 0)$, $P(X \leq 4)$, $P(X > 3)$.
    e. $X \sim \mathcal{P}(5)$ izanik: $P(X = 8)$, $P(X \leq 8)$, $P(5 \leq X \leq 8)$, $P(X > 2)$.

```
# Rko funtzioak erabili ditzakezu banaketa ezagunen
# probabilitateak kalkulatzeko. Adibidez:
# X ~ Bin(10, 0.3) izanik, P(X <= 2)
pbinom(2, size=10, p=0.3) # Banaketa-funtzioa
[1] 0.3827828

# Laguntzak
?pnbinom # Binomial negatiboarena
?ppois # Poissonen banaketarena
?phyper # Hipergeometrikoa
```

18. Produkzio-kate batean piezak 8ko sortetan egiten dira. Piezak A ala B motakoak izan daitezke zoriz. Pieza bat A motakoa izateko probabilitatea 0.6 da. Gainera, pieza bakoitzaren mota sortako gainontzeko piezen motarekiko askea dela suposaturik, kalkula itzazu:

    a. Sortako pieza guztiak mota berekoak izateko probabilitatea.
    b. Lehenengo hirurak, mota batekoak eta gaineratikoak bestekoak izateko probabilitatea.
    c. A motako 5 pieza eta B motako 3 pieza izateko probabilitatea.
    d. Gutxienez A motako bost pieza izateko probabilitatea.
    e. Lehenengo biak A motakoak eta gutxienez A motako bost pieza izateko probabilitatea.

19. CD-irakurgailuetako zirkuito-txartelen % 5ek arazoak ematen ditu 3 hilabeteko epean. Hori guztia ikertzeko 25 irakurgailutako lagina aukeratu da. Kalkula itzazu:

    a. Laginean gehienez bi irakurgailuk ematea arazoak.
    b. Laginean arazoak ematen dituen irakurgailurik ez topatzea.
    c. Oro har, laginean arazoak ematen dituen zenbat irakurgailu egotea espero behar da?

20. Hiri jakin bateko udalak egindako azterketa batean, honako hau ondorioztatu da: autoa bidegurutze batera iritsi, semaforoa gorri aurkitu eta beste autorik

bistan izan ezean, autoen % 20k aurrera jarraitzen du semaforo gorriari kasurik egin gabe. Baldintza horietan aurkitutako 9 autoko lagina hartuz, kalkulatu:

a. Gehienez 3 auto gelditzeko probabilitatea.
b. Zehazki 3 auto gelditzeko probabilitatea.
c. Gutxienez 3 auto gelditzeko probabilitatea.
d. Zenbat auto gelditzea espero behar da?

21. Demagun hipoteka bat eskatzeko egin beharreko paperak % 45ek ez dituela ongi betetzen lehenengo aldian. Eskaria egin dutenen artean 10 lagun aukeratu ditugu. Kalkula itzazu:

a. Gutxienez 2 lagunek lehenengo aldian ongi ez betetzeko probabilitatea.
b. Lehenengo aldian paperak ongi bete gabeko lagunak 1 eta 3 bitartean izateko probabilitatea.
c. Gehienez bi lagunek ongi betetzeko probabilitatea.

22. Tabernek ke-detektagailuak instalaturik izatea oso komenigarria dela dio suhiltzaile-departamentuak. Hori dela eta, ikuskapen orokor bat martxan jartzeko intentzioa du, baldin eta ke-detektagailuak dituzten tabernen benetako portzentaia (p) % 80 baino nabarmenki txikiagoa bada. Hori horrela izan daitekeen ala ez ikusteko, 10 taberna zoriz aukeratu eta ke-detektagailua zenbatek duten zenbatuko du. Demagun ke-detektagailuak 6 tabernetan edo gutxiagotan bakarrik aurkituz gero, martxan jarriko duela ikuskapen orokorra; bestela ez. Orduan:

a. $p = 0.8$ izango balitz, zein da ikuskapen orokorra martxan ez jartzeko probabilitatea?
b. $p = 0.7$ izango balitz, zein da ikuskapen orokorra martxan ez jartzeko probabilitatea? Eta $p = 0.6$ balitz?
c. Aurreko ataletako emaitzak ikusi ondoren, nolakoa iruditzen zaizu erabakia hartzeko irizpidea?

23. Populazioko gizabanakoen % 0.003a urtero gaixotasun arraroren baten ondorioz hiltzen dela ezaguna da osasun-sailean. Kalkula itzazu:

a. Aurrez finkaturiko urtean, 10000 biztanleren artetik gaixotasun arraro horren ondorioz hildakoak 2 baino gehiago izateko probabilitatea.
b. Gaixotasun arraro horren ondorioz espero izateko hildako kopurua.

24. Geologo batek basalto-harkaitz 10 zati eta granito-harkaitz beste 10 zati bildu ditu. Lan baterako 15 harkaitz zati aukeratu behar direla jakinik:

a. Zein probabilitate-lege izango du aukeratuen artean topatutako basaltozko zati kopurua adierazten duen zorizko aldagaiak?
b. Zein da bai basaltozko edo bai granitozko zati guztiak aukeratuak izateko probabilitatea?

25. Animali-kolonia batean, aleak ez badu medikamenturik jaso, alearen ondorengo kopurua 6 parametroko Poissonen banaketaren bidez ongi deskribatua dagoen z.a. da. Baina aleari medikamentua eman bazaio, orduan, beraren ondorengo kopurua 3 parametroko Poissonen banaketa jarraitzen duen z.a. da. Koloniaren % 30a medikatua dagoela suposatzen da.

    a. Zein da zoriz aukeraturiko aleak gehinez 5 ondorengo izateko probabilitatea?
    b. Aleak 5 ondorengo izan dituela jakinik, zein da medikamentu-pean egoteko probabilitatea?
    c. Kolonian, zein da aleko batez besteko ondorengo kopurua?

26. Medikuntza-ikerketa batek dioenez, 200 pertsonatik batek hesteko minbiziaren gena darama. Demagun 999 pertsonako lagin bat hartzen dugula. Zein izan daitekeela uste duzu laginean gen hori daramatenen kopurua adierazten duen zorizko aldagaiaren gutxi gorabeherako banaketa? Horren baitan,

    a. Laginean 4 eta 7 (biak barne) pertsonek gen hori izatea.
    b. Gutxienez lagineko 8 pertsonek gen hori izatea.

27. Zirkuitu-txartelen froga bat egiten ari da enpresa jakin bat, eta badakite txarteleko diodo batek huts egiteko probabilitatea 0.01 dela. Txartel batek 200 diodo dituela jakinik:

    a. Txartel batean zenbat diodok huts egitea espero da? Zein da huts egingo duen diodo kopuruaren bariantza?
    b. Zein da zoriz aukeratutako txartelean gutxienez 4 diodok huts egiteko probabilitatea?
    c. Bezero bati 5 txartel saldu zaizkio. Txartel bakoitzaren funtzionamendua besteekiko askea dela jakinik eta txartelak ongi funtzionatzeko diodo guztiek ongi funtzionatu behar dutela kontutan hartuz, zein da gutxienez 4 txartelek ongi funtzionatzeko probabilitatea?

28. Izan bedi:

$$f(x) = \begin{cases} \frac{1}{2}e^{-\lambda}\frac{\lambda^x}{x!} + \frac{1}{2}e^{-\mu}\frac{\mu^x}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{bestela} \end{cases}$$

    a. Froga ezazu funtzio hori probabilitate-legea dela.
    b. Demagun enpresa batek bi itzultzaile dituela. Lehenengoak batez beste orrialdeko λ akats egiten ditu, eta bigarrenak, berriz, orrialdeko μ akats. Enpresan egin beharreko itzulpen guztiak erdibana egiten badituzte eta X z.a.-k orrialde batean topatutako akats kopurua adierazten badu, dagokion probabilitate-legea goiko aldean aipaturiko $f(x)$ egokia izan daiteke. Zein izango litzateke X-ren probabilitate-legea baldin eta lehen itzultzaileak itzulpen guztien % 60a itzultzen badu?
    c. Zein da EX?
    d. Eta VAR(X)?

Aldagai jarraituak

1. Birus mota baten bizitza (ordutan) honako dentsitate-funtzio hau duen zorizko aldagaia da:

$$f(x) = \begin{cases} \dfrac{k}{x^4}, & x \geq 1 \\ \quad , & \\ 0\ , & x < 1 \end{cases}$$

   a. Kalkula ezazu $f(x)$ dentsitate-funtzioa izateko k-ren balioa.
   b. Zein da birus mota horien batez besteko bizitza?
   c. Aurkitu ezazu banaketa-funtzioa.
   d. Kalkula ezazu desbideratze estandarra.
   e. Aurkitu ezazu zoriz aukeraturiko birusa 5 ordu baino gehiago bizitzeko probabilitatea.

2. Ospitale jakin batean frogatu da honako dentsitate-funtzio hau duen zorizko aldagaia egokia dela umeen jaiotze-pisua (kilotan) adierazteko:

$$f(x) = \begin{cases} kx, & 2 \leq x \leq 4 \\ 0\ , & \text{beste kasuetan} \end{cases}$$

   Horren baitan:

   a. Aurkitu ezazu k konstantearen balioa $f(x)$ dentsitate-funtzioa izan dadin.
   b. Aurkitu ezazu banaketa-funtzioa.
   c. Kalkula itzazu X-ren itxaropena, bariantza eta desbideratze estandarra.
   d. Kalkula ezazu zoriz aukeratutako ume baten pisua 3 kg baino handiagoa izateko probabilitatea.
   e. Zenbat kilo izan beharko ditu ume batek, umeen % 90aren pisua balio horretatik beherakoa izan dadin?

3. Prozesu kimiko batek erreakzionatzeko tenperaturak ($^{\circ}$C) $a = -5$ eta $b = 5$ parametrodun banaketa uniformeari jarraitzen dio.

   a. Irudika ezazu dentsitate-funtzioa.
   b. Kalkula itzazu: $P(X < 0)$, $P(-2 < X < 2)$, $P(-2 \leq X \leq 3)$.
   c. $P(k < X < k+4)$, k konstanteak $-5 < k$ eta $k+4 < 5$ betetzen dituela kontutan hartuz.

4. Familietako sarrera kopuruaren (X) banaketa hurbiltzeko, askotan egokia izaten da Paretoren banaketa izenekoa. Banaketa horrek bi parametro ditu, k eta θ, biak positiboak, eta hau da dagokion dentsitate-funtzioa:

$$f(x) = \begin{cases} \dfrac{k\theta^k}{x^{k+1}}, & x \geq \theta \\ 0\ , & \text{beste kasuetan} \end{cases}$$

a. Irudika ezazu dentsitate-funtzioaren grafikoa.
b. Egiazta ezazu dentsitate-funtzioaren azpialdeko azalera 1 dela.
c. $x_0 > \theta$ balioa finkaturik, kalkula ezazu $P(X < x_0)$-rentzako espresioa.

5. X zorizko aldagai baten dentsitate-funtzioa honako hau da:

$$f(x) = \begin{cases} k\,(1-(x-3)^2), & 2 \le x \le 4 \\ 0 \qquad\quad , & \text{bestela} \end{cases}$$

a. Benetan dentsitate-funtzioa izateko, zein balio du k konstanteak?
b. Irudika ezazu dentsitate-funtzioa.
c. Irudian oinarrituz, zein da E(X)ren gutxi gorabeherako balioa?

6. Londres-Sydney hegaldia oro har bere orduan iristen bada ere, zenbaitetan aurreikusitako orduan baino lehenago ala beranduango irits daiteke. Izan bedi X zorizko aldagaia, hegaldiaren atzerapena (ordutan) adierazten duena honako dentsitate-funtzio hau duelarik:

$$f(x) = \begin{cases} e^{-(x+1)}\,, & x \ge -1 \\ 0\,, & x < -1 \end{cases}$$

a. Kalkula ezazu bere banaketa-funtzioa.

b. 0.9 probabilitatearekin, atzerapena asko jota zenbatekoa izango da eredu horren arabera?

c. Konpainiak atzerapenaren araberako kalte-ordainak ezarri nahi ditu honela:

- 100 euroko kalte-ordaina, ordubeteko atzerapena baino gehiagokoa izan bada baina 2 ordu baino gutxiagokoa.
- 120 euroko kalte-ordaina, 2 orduko atzerapena baino gehiagokoa izan bada baina 5 ordu baino gutxiagokoa.
- 300 euroko kalte-ordaina, 5 orduko atzerapena baino gehiagokoa izan bada.

Kalte-ordainak horrela ezartzen baditu, batez beste hegaldiko zenbateko kalte-ordaina ordaintzea espero du?

7. Enpresa banatzaile batek asteko jasotzen duen propano-gasaren eskaeraren (milaka galoitan) dentsitate-funtzioa honoko hau da:

$$f(x) = \begin{cases} 2\,(1-1/x^2), & 1 \le x \le 2 \\ 0 \qquad\quad , & \text{bestela} \end{cases}$$

a. Kalkula ezazu banaketa-funtzioa.
b. Zein da medianaren balioa?
c. Kalkula itzazu $E(X)$ eta VAR(X).
d. Baldin eta enpresa honetan aste hasieran 1500 galoi-propano badago eta astean zehar ez badu gehiagorik jasoko, astea amaitzean zenbat propano gelditzea espero da? Enpresa gai izango al da duen propano-eskaerari aurre egiteko?

8. Demagun $Z \sim N(0,1)$ dela. Pentsa ezazu egoeraren bat, non arrazonagarria den Z zorizko aldagaiak $N(0,1)$ banaketari jarrai diezaiokeela. Kalkula itzazu honako probabilitate hauek eta irudikatu probabilitate bakoitzari dagokion azalera.

   a) $P(Z > 2)$
   b) $P(Z \geq -0.97)$
   c) $P(0 \leq Z \leq 2.17)$
   d) $P(-2.5 \leq Z \leq 0)$
   e) $P(-2.5 \leq Z \leq 2.5)$
   f) $P(Z < 1.37)$
   g) $P(|Z| \leq 2.5)$
   h) $P(|Z| > 2.5)$
   i) $P(Z > z_0) = 0.03 \Rightarrow z_0 = ?$
   j) $P(Z \leq z_0) = 0.0025 \Rightarrow z_0 = ?$
   k) $P(Z \geq z_0) = 0.98 \Rightarrow z_0 = ?$

```
# Rko funtzioak erabili ditzakezu banaketa ezagunen
# probabilitateak kalkulatzeko. Adibidez:
# Z ~ N(0, 1) izanik, P(Z <= 2)
pnorm(2) # Banaketa-funtzioa
[1] 0.9772499

# z0 balioa bilatzeko, P(Z <= z0)=0.90 delarik
qnorm(0.90)
[1] 1.281552
```

9. Demagun $X \sim N(52,36)$ dela. Kalkula itzazu honako probabilitate hauek eta irudikatu probabilitate bakoitzari dagokion azalera.

   a) $P(X > 50)$
   b) $P(X \leq 55)$
   c) $P(48 \leq X \leq 56)$
   d) $P(|X - 52| \leq 6)$
   e) $P(X > x_0) = 0.24 \Rightarrow x_0 = ?$
   f) $P(X \leq x_0) = 0.175 \Rightarrow x_0 = ?$

```
# Rko funtzioak erabili ditzakezu banaketa ezagunen
# probabilitateak kalkulatzeko. Adibidez:
# X ~ N(52, 36) izanik, P(X <= 48)
pnorm(48, mean=52, sd=6) # Banaketa-funtzioa
[1] 0.2524925

# x0 balioa bilatzeko, P(X <=x0)=0.90 delarik
qnorm(0.90, mean=52, sd=6)
[1] 59.68931
```

10. Portatilen baterien biziraupen denbora (urtetan) adierazten duen zorizko aldagaiak hoako dentsitate-funtzio hau du:

$$f(x) = \begin{cases} -\frac{x}{2} + 1, & 0 \leq x \leq 2 \\ 0 \quad , & \text{bestela} \end{cases}$$

   a. Aurkitu ezazu banaketa-funtzioa.

b. Kalkula itzazu $P(X \geq 1)$, $P(X > 1)$ eta $P(X = 1)$.
c. Kalkula ezazu bateria horiek espero izateko biziraupen-denbora.
d. Horrelako 100 bateria zoriz aukeratzen dira. Zein da horietatik 10 baino gehiagok gutxienez 6 hilabetez bizirauteko probabilitatea?

11. Auto-kontzesionario batean marka ezagun bateko autoak saltzen dira. Horrelako auto bat, bi urte pasatu ondoren, martxan egoteko probabilitatea 0.8 dela eta saldutako auto kopurua 4000 izan dela jakinik, kalkula ezazu:

    a. Bi urte pasatu ondoren, 3200 auto baino gehiago martxan egoteko probabilitatea.
    b. Bi urte pasatu ondoren, 3150 auto baino gehiago martxan egoteko probabilitatea.

12. Motorreko zilindroak egiteko makina jakin bat erabiltzen da, eta badakigu zilindro horien diametroak banaketa normalari jarraitzen diola. Zein da zoriz aukeraturiko zilindro bat honako tarte hauetan egoteko probabilitatea?

    a. Batezbestekotik 1.5σ distantziara edo gutxiagora?
    b. Batezbestekotik 2.5σ distantziara edo gehiagora?
    c. Batezbestekotik 1σ eta 2σ distantziarako tartean?

13. Ardo-botiletan erabiltzeko kortxoak egiten dituzten bi makina ditugu. Lehen makinak egiten dituen kortxoen diametroak 3 cm-ko batezbestekoa eta 0.1 cm-ko desbideratze estandarreko banaketa normalari jarraitzen dio. Bigarren makinak egiten dituenen diametroak, aldiz 3.04 cm-ko batezbestekoa eta 0.02 cm-ko desbideratze estandarreko banaketa normalari jarraitzen dio. Baleko kortxoak diametroa 2.9 cm eta 3.1 cm artean dutenak direla jakinik, zein makinak du baleko kortxo bat egiteko probabilitate handiagoa?

14. Izan bitez X eta Y zorizko aldagaiak, $Y = aX + b$ izanik ($a, b \in \mathbb{R}$ eta $a \neq 0$).

    a. Froga ezazu:
    $$X \sim N(\mu, \sigma^2) \Rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$$
    Iradokizuna: idatz ezazu Y-ren banaketa-funtzioa, $P(Y \leq y)$, X-ren dentsitate-funtzioa agertzen den moduan.
    b. Pintura-potoen fabrikazioan ikusi da potoen erradioak 10 cm-ko batezbestekoa eta 0.1 cm-ko desbideratze estandarra dituen banaketa normalari jarraitzen diola. Zein da potoaren tapa eusteko erabiliko den zinta itsasgarriak 62 cm baino txikiagoa izateko probabilitatea?

15. Metalen Rockweel gogortasuna neurtzeko, lehenengo, metala altzairuzko puntadun tresna batekin kolpatu eta, ondoren, sortutako urraduren sakonera neurtzen da. Demagun aleazio jakin baten gogortasunak 70 batezbesteko eta 3 desbideratze estandarreko banaketa normalari jarraitzen diola. Metal hori zenbait pieza fabrikatzeko erabiliko da.

    a. Baldin eta aleazio horrekin egingo diren piezak balekoak izateko Rockweek gogortasunak 67 eta 75 artean egon behar badu, zein da zoriz

aukeraturiko pieza bat balekoa izateko probabilitatea?

b. $(70-c, 70+c)$ itxurako zein tartetan egongo da piezen % 95a?

c. Zoriz 10 pieza aukeratu dira. Horietatik zenbat espero da balekoak izatea?

16. Kalkula itzazu honako probabilitate hauek:

$a)\ P(\chi^2_{10} > 12.55)$
$b)\ P(\chi_{16} < 28.85)$
$c)\ P(\chi^2_{21} > a) = 0.05 \Rightarrow a = ?$
$d)\ P(\chi^2_{30} \leq a) = 0.90 \Rightarrow a = ?$
$e)\ P(\chi^2_{50} \leq a) = 0.975 \Rightarrow a = ?$
$f)\ P(a_1 \leq \chi^2_{25} \leq a_2) = 0.95 \Rightarrow a_1, a_2$ $P(\chi^2_{25} < a_1) = P(\chi^2_{25} > a_2)$izanik
$g)\ P(t_{10} \leq 2.764)$
$h)\ P(t_8 \leq -1.397)$
$i)\ P(t_{10} \leq a) = 0.90 \Rightarrow a = ?$
$j)\ P(t_{30} > a) = 0.025 \Rightarrow a = ?$
$k)\ P(-a \leq t_{50} \leq a) = 0.95 \Rightarrow a = ?$
$l)\ P(F_{24,30} > 1.89)$

```
# Rko funtzioak erabili ditzakezu. Adibidez:
# X ~ KI^2, 10 askatasun gradurekin, izanik, P(X <= 12.55)
 pchisq(12.55, df=10) # Banaketa-funtzioa
[1] 0.7500693

# x0 balioa bilatzeko, P(X <=x0)=0.90 delarik
qchisq(0.90, df=10)
[1] 15.98718

# Modu beretsuan,
# pt() eta qt() funtzioak erabil ditzakezu Student-t banaketarako
```

17. Fabrikatze-prozesu batean agertzen diren eguneroko akastun unitate kopurua 10eko itxaropena eta 9 bariantza dauzkan zorizko aldagaia da. Determina ezazu 150 egunetan sortutako akastun unitate kopuruak 1480 unitate gainditzeko probabilitatea.

18. Produkzio-katean egiten diren piezen artean, badakite akastun kopuruak 0.5 parametrodun Poissonen banaketari jarraitzen diola. Gainera, katea egunero jartzen da martxan, eta egindako piezak jaso egiten dira.

    a. Zein da hiruhileko batean espero duten akastun kopurua?

    b. Zein da hiruhilekoan zehar gehienez 60 akastun ekoizteko probabilitatea? Eta gehienez 40 akastun ekoiztekoa?

19. Denda batean saltzen den artikulu baten salneurria p da. Egun batean zehar denda horretan saltzen den artikulu kopuru osoa (denetarikoa) 100 da, eta horietatik % 30a p salneurriko artikuluari dagokio. Kalkula ezazu:

    a. Saldutako artikuluetatik 40 baino gehiago p salneurrikoak izateko probabilitatea.

b. Demagun bi hilabetean zehar (60 egun) salmentaren azterketa bat egiten dela. Eguneroko salmentaren independentzia kontuan izanik, kalkula ezazu 60 egun horietan saldutako p salneurriko artikulu kopurua 1700 baino gehiagokoa ez izateko probabilitatea.

20. Demagun X zorizko aldagaiak osagai hidrauliko baten lehen konponketa behar izan arteko iraupena (urtetan) adierazten duela eta hau dela bere dentsitate-funtzioa:

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0 \\ 0 \quad , & x \leq 0 \end{cases}$$

a. Froga ezazu benetan dentsitate-funtzioa dela.
b. Kalkula ezazu bere banaketa-funtzioa.
c. Erabil ezazu aurreko ataleko emaitza, 2 eta 5 urte bitartean lehen konponketa behar izateko probabilitatea kalkulatzeko.
d. x urtera konponketa behar duen osagai bakoitzari «berreskuratze-indize» bat egokitzen zaio $\frac{100}{x+4}$ erara. Zein da osagaien espero izateko berreskuratze-indizea?

21. X zorizko aldagaia honela definituta dago:

$$f(x) = \begin{cases} 20x^3(1-x), & 0 \leq x \leq 1 \\ 0 \quad , & \text{bestela} \end{cases}$$

a. Egiaztatu dentsitate-funtzioa dela.
b. Kalkulatu dagokion banaketa-funtzioa.
c. Kalkulatu $P(X \leq 2/3)$.
d. Zorizko beste aldagaia definitzen da: baldin eta $1/3 < X < 2/3$ bada, orduan $K = k_1$, bestela $K = k_2$. Kalkulatu K zorizko aldagaiaren espero izateko balioa.

22. Demagun X aldagaiak adierazten duela ondorio larriak jasan baino lehen gizabanakoak bereganatu dezakeen irradiazio kantitatea (roentgen unitatean). X zorizko aldagaia normala dela onartzen da, bere itxaropena 500 roentgen eta desbideratze estandarra 150 roentgen direlarik.

a. Zein kantitatek beteko du, non hortik aurrera % 5ek bakarrik ez duela jasango ondorio larririk?

b. Kalkula ezazu: $P(350 < X < 650)$.

c. 100 gizabanako aukeratzen ditugu. Demagun bakoitzak bereganatu dezakeen irradiazio kantitatea askea dela. Kalkula itzazu:

- Horien artean, gutxienez 82 gizabanakok 350 eta 650 roentgen bitarteko irradiazio kantitatea bereganatzeko probabilitatea.
- 100 gizabanakoek berengana dezaketen irradiazio kantitate osoa 54000 roentgen baino handiagoa izateko probabilitatea.

23. Programazioko ikastaro bateko lehen praktikan, programa erraz bat idatzi eta exekutatzeko eskatzen zaie ikasleei. Ikasleen % 40k arazorik gabe burutzen duela jakinik, kalkula itzazu honako probabilitate hauek 50 laguneko ikasgela baterako:

    a. Gutxienez 25 ikaslek praktika arazorik gabe burutzea.
    b. Arazorik gabe praktika burututako ikasle kopurua 15 eta 25 bitartean egotea.

24. Ordenadoretarako soinu-txartel eredu berri baten ekoizpenean buru-belarri dabil enpresa bat. Ekoizpen horretan, badakite astean zehar izandako soinu-txartel akastunen kopuruak 1 parametrodun Poissonen banaketari jarraitzen diola.

    a. Zein da urtean zehar (52 aste) espero duten txarte kopuru akastuna?
    b. Zein da urtean zehar gehienez 38 akastun ekoizteko probabilitatea?

25. Lurralde bateko biztanleek A Gobernuz Kanpoko Erakundeari urtean egindako ematea uniformeki banatzen da 5 euro eta 50 euro bitartean. Kalkula ezazu zoriz aukeratutako 150 pertsonek egindako emateak guztira 2500 euro gainditzeko probabilitatea.

8. gaia

INFERENTZIA ESTATISTIKOA. Estimazioa.

1. Fabrika batean ekoitzitako piezek ez lukete inolako urraturik izan behar, baina jakin badakite zenbait piezek urratuak izaten dituztela. Piezak zenbateraino urratuak agertzen diren aztertu asmoz, 150 piezen urratu kopurua zenbatu dute, eta hauek dira lortutako emaitzak:

Piezako urratu kopurua	0	1	2	3	4	5	6	7
Behatutako maiztasuna	18	37	42	30	13	7	2	1

    Izan bedi $X \equiv$ *«Piezaren urratu-kopurua»* adierazten duen zorizko aldagaia, eta onar dezagun λ parametrodun Poissonen banaketari jarraitzen diola.

    a. Aurkitu ezazu λ-rentzako estimatzaile zentratua eta behatutako datuen araberako estimazioa. (*Iradokizuna:* gogoratu $EX = \lambda$ dela, eta $E(\bar{X}) = ?$)
    b. Zein da aurkitu duzun estimatzailearen desbideratze estandarra (errore estandarra)? Kalkula ezazu errore estandarraren estimazioa.

2. Sistema informatiko batean erabiltzen diren osagai mota bateko 80 osagaiko zorizko lagin batean, 12 akastunak zirela ikusi zen.

    a. Eman ezazu akatsik gabeko osagaien proportzioaren estimazioa.

b. Demagun osagai horiek binaka seriean konektatutako sistemak osatuko ditugula. Sistemak ongi funtzionatzeko osagai bakoitzak ongi funtzionatu behar duela kontuan izanik, eman ezazu akastun sistemen proportzioaren estimazioa.

3. Ogi mota bateko (A mota) 12 aledun zorizko lagin batean, ogiaren karbono-hidrato edukia aztertu zen, eta honako emaitza hauek lortu ziren: 76.93, 76.88, 77.07, 76.68, 76.39, 75.09, 76.88, 77.67, 78.15, 76.50, 77.16, 76.42. Karbono-hidrato edukiaren banaketa normala dela jakinik,

   a. Estima ezazu benetako karbono-hidrato edukiaren batezbestekoa (μ_X).
   b. Zein da estimatzaile horren errore estandarra?
   c. Beste ogi mota baten (B mota) karbono-hidrato edukia ere aztertu nahi da. Demagun A motako karbono-hidrato edukiari dagokion lagina $X_1, X_2, ..., X_n$ moduan adieraziko dugula, eta B motakoari dagokiona, berriz, $Y_1, Y_2, ..., Y_m$ moduan. Froga ezazu $\bar{X} - \bar{Y}$ estimatzailea zentratua dela $(\mu_X - \mu_Y)$-rentzat.
   d. Baldin eta B motako ogiarentzat 14 tamainako lagina behatu bada eta $\bar{y} = 74.28$ kalkulatu, eman ezazu $(\mu_X - \mu_Y)$-rentzat puntu-estimazio bat.
   e. Zein da (c) ataleko estimatzailearen errore estandarra? Baldin eta $s_X = 0.75$ eta $s_Y = 0.70$ izan badira, eman ezazu errore estandarraren estimazioa.

4. Demagun μ luzerako hagaxkekin karratuak eraikiko direla. Beraz, karratuaren azalera printzipioz μ^2 izango da. μ balioa ezezaguna denez, $X_1, ..., X_n$, hagaxken luzerak adierazten duten zorizko lagina hartuko da. X_i, zorizko aldagai bakoitzak μ espero izateko balioa eta σ^2 bariantza dituela kontsideratuz:

   a. Froga ezazu $\bar{X}^2$ estimatzaile alboratua dela μ^2-rentzat. (*Iradok*izuna: gogoratu edozein Y zorizko aldagairentzat $E(Y^2) = VAR(Y) + E(Y)^2$ betetzen dela. Aplikatu hori $Y = \bar{X}$-rekin).
   b. k-ren zein baliotarako izango da $\bar{X}^2 - kS^2$ estimatzailea μ^2-ren estimatzaile alboragabea?

5. Komunikazio-protokoloetarako bi parametrizazio alderatu nahi dira, A eta B. A parametrizazioa erabiliz, bidalitako paketeen artean n_1 zoriz aukeratu eta X_1-ek ez du kanpoko erasoen erasanik izan. B parametrizazioa erabiliz, bidalitako paketeen artean n_2 zoriz aukeratu eta X_2-k ez du kanpoko erasoen erasanik izan. Izenda ditzagun p_1 eta p_2 moduan, A eta B parametrizazioak erabiliz, hurrenez hurren, kanpoko erasoen erasanik jasaten ez duten bidalitako paketeen portzentajeak.

   a. $\frac{X_1}{n_1} - \frac{X_2}{n_2}$ estimatzailea alboragabea al da $(p_1 - p_2)$-rentzat? (*Iradokizuna:* erreparatu zein den X_i bakoitzaren banaketa eta, ondorioz, zein den bere itxaropen matematikoa).

b. Zein da a) ataleko estimatzailearen errore estandarra?
c. Baldin eta $n_1 = n_2 = 200$ bada eta behatutako balioak $x_1 = 127$ eta $x_2 = 176$ izan badira, erabil ezazu a) ataleko estimatzailea $(p_1 - p_2)$-ren estimazio bat lortzeko. Zein errore estandar estimatzen da?

6. Demagun honako dentsitate-funtzio hau duen populaziotik $X_1, X_2, ..., X_n$ zorizko lagina atera dugula.

$$f(x) = \frac{1}{2}(1 + \theta x), \quad -1 \leq x \leq 1 \quad (-1 \leq \theta \leq 1 \text{ izanik}).$$

Froga ezazu $\hat{\theta} = 3\bar{X}$ estimatzailea alboragabea dela θ-rentzat. *Iradokizuna:* lehenengo, kalkula ezazu $\bar{X}$-ren itxaropen matematikoa.)

7. Demagun Rayleigh banaketa jarraitzen duen populaziotik $X_1, X_2, ..., X_n$ zorizko lagina atera dugula. Rayleigh banaketaren dentsitate-funtzioa: $f(x) = \frac{x}{\theta} \exp\left\{\frac{-x^2}{2\theta}\right\}, \quad x > 0.$

a. Ikus daiteke $E(X^2) = 2\theta$ dela. Emaitza hori erabiliz eta $\sum X_i^2$ gaian oinarrituz, aurkitu ezazu θ-rentzat estimatzaile alboragabe bat.
b. Erabili ezazu aurreko ataleko estimatzailea honako lagin honetatik lortu dezakegun θ-ren estimazioa kalkulatzeko.

Lagina: 16.88, 10.23, 4.59, 6.66, 13.68, 14.23, 19.87, 9.40, 6.51, 10.95

8. A motako ongarriarekin μ_1 batezbestekoa eta σ^2 bariantzako produkzioa lortzen da, eta B motakoarekin, berriz, μ_2 batezbestekoa eta σ^2 bariantzakoa. A motako ongarriarekin lortutako produkzioaren n_1 tamainako eta B motakoarekin lortutakoaren n_2 tamainako zorizko laginak ateraz gero, adieraz ditzagun S_1^2 eta S_2^2 bidez lagin bakoitzeko kuasibariantzak, hurrenez hurren. Froga ezazu honako estimatzaile hau σ^2-rentzat estimatzaile alboragabea dela.

$$\hat{\sigma^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

9. Demagun 1000 zorizko lagin desberdin ditugula eta lagin guztiak modu askean aukeratu direla. Datu horietatik, μ, populazioko benetako batez besteko baliorako % 95eko konfiantza-tarteak kalkulatu dira.

a. Horietako zenbat tartek barneratuko dutela μ balioa espero da?
b. Kalkula ezazu μ barneratzen duten tarteen kopurua 940 eta 960 kopuruen artekoa izateko probabilitatea. (*Iradokizuna:* izan bedi $Y \equiv$ *«1000 konfiantza-tarteen artean μ barneratzen dutenen kopurua»*. Zein banaketa du Y-k?)

10. Jakina da informatikariak karrera bukatu eta lana bilatzeko orduan oso komenigarria dela A lengoaia jakitea. Horregatik, A lengoaia irakasteko ikastaro bat antolatzeko asmoa du informatikari talde batek, baldin eta lengoaia

hori menperatzen ez dutenen proportzioa altua bada. Proportzio hori estimatzeko, lagin bat bildu eta 250 informatikaritatik 113k ez dutela menperatzen ikusi dute. Kalkula ezazu proportzio horrentzako %95 eko konfiantza-tartea.

11. Lurralde jakin batean behatu diren 110 tximistek radarretan izandako oihartzunaren iraupena 0.81 segundoko batezbestekoa eta 0.34 segundoko desbideratze estandarra izan dutela behatu da. Kalkula ezazu benetako oihartzunaren iraupenaren batez bestekorako % 99ko konfiantza-tartea.

12. 40 transistoretako lagin batean behatutako asetasun-korrontearen lagin-batezbestekoa 9.46 miliampere izan da. Lagin-desbideratze estandarra, berriz, 0.58 miliampere. Kalkula ezazu populazioko batez besteko asetasun-korrontearentzako % 95eko konfiantza-tartea. Tarte hori kalkulatzeko suposizioaren bat onartu behar izan al duzu?

13. Banku-etxe bateko finantza-arduradunak honako informazio hau bildu du: zoriz aukeraturiko 200 bezerok erosketak kreditu-txartelarekin ordaindu ditu, baina horietatik 136k beren kontu korronteko saldo eza dela-eta interesak ordaindu behar izan dituzte.

    a. Kalkula ezazu kreditu-txartela erabili eta interesak ordaindu behar izaten dituztenen portzentajearen puntu-estimazioa eta % 90eko konfiantza-tartea.
    b. Aurreko ataleko informazioa erabiliz, % 90eko konfiantza-tarteak 0.05 luzera baino gehiago izaterik ez badugu nahi, zein lagin-tamaina aukeratu behar dugu?

14. Suhiltzaileek erabiltzen dituzten maskaren kalitatea bermatzeko, 35 maskara 125 ºC-ko tenperaturapean jarri dira. Horietako 11ak, ordea, kalteak jasan dituzte tenperatura altua dela eta. Kalkula ezazu 125 ºC-ra kalteak jasaten dituzten maskaren proportziorako % 90eko konfiantza-tartea.

15. Jakina da amiantoaren eraginpean luzaro egotea osasunarentzat oso kaltegarria dela. Berriki egindako ikerketa batean, amiantoaren eraginpean luzaro egondako 16 langileren osasun-egoera aztertu da. Langile horien biriken «gaitasuna» neurtu (cm^3/cmH_2O) eta honako emaitza hauek behatu dira:

167.9	180.8	184.8	189.8	194.8	200.2	201.9	206.9
207.2	208.4	226.3	227.7	228.5	232.4	239.8	258.6

    Baldin eta biriken gaitasunak banaketa normalari jarraitzen badio, kalkula ezazu batez besteko biriken gaitasunerako % 95eko konfiantza-tartea.

16. Nikel eduki jakin bat duen altzairuaren hausturarako erresistentzia behatu da zorizko lagin batean, eta hauek dira lortutako emaitzak:

69.5	71.9	72.6	73.1	73.3	73.5	75.5	75.7	75.8	76.1	76.2

76.2	77.0	77.9	78.1	79.6	79.7	79.9	80.1	82.2	83.7	93.7

Kalkula ezazu erresistentziaren bariantza eta desbideratze estandarrerako % 99ko konfiantza-tartea. Erresistentziaren banaketa edozein izanda ere, lortutako tartea balekoa al da?

17. Iaz mintegiko bi mila milioi landarerekin 3 milioi akre baso baino gehiago berritu ziren. Udan egon zen izugarrizko lehortea dela eta, horietariko landare asko hil egin ziren ordea. 1000 landarek osoturiko lagina hartu zen, eta hauetatik 300 hilda zeuden. Kalkula ezazu hildako mintegiko landare-proportziorako % 90eko konfiantza-tartea. Erabil ezazu informazio hori, populazioan hildako mintegiko landare kopuru osoa estimatzeko.

18. Ahotsa automatikoki ezagutzeko sistema baten errendimendua neurtu nahi da. Horretarako, 40 hitz zoriz aukeratu dira, eta 30 ongi ezagutu ditu sistemak.

    a. Aurkitu ezazu p, sistemak ongi ezagututako hitzen proportziaren puntu-estimazioa.
    b. Aurkitu ezazu sisteman ongi ezagututako hitzen proportziorako % 95eko konfiantza-tartea. Tarte horretan oinarrituz, p-ren balioaren ideia ona izateko nahiko handia da n, laginaren tamaina?
    c. Zein lagin-tamaina beharko genuke 0.02ko distantziara eta % 95eko konfiantza-mailaz ongi ezagututako hitzen proportzioa estimatzeko?

9. gaia

INFERENTZIA ESTATISTIKOA. Hipotesi-probak.

1. Honako p-balio hauek emanik, zein kasutan baztertuko genuke hipotesi nulua 0.05 mailako froga bat egindakoan?

    a. 0.001
    b. 0.021
    c. 0.078
    d. 0.047
    e. 0.148

2. Esan ezazu bikote bakoitzaren kasuan behatutako p-balioak H_0 baztertzera behartzen duen ala ez, emandako α adierazgarritasun-maila kontuan hartuta.

    a. $p = 0.084$, $\alpha = 0.05$
    b. $p = 0.003$, $\alpha = 0.001$
    c. $p = 0.498$, $\alpha = 0.05$
    d. $p = 0.084$, $\alpha = 0.10$

e. $p = 0.039$, $\alpha = 0.01$
f. $p = 0.218$, $\alpha = 0.10$

3. Automobil jakin baten gurpilak, erosi berriak direnean, 30 lb/plg^2-ko presioz bete beharko lirateke. Presioaren benetako batezbestekoa μ izanik, aurkitu ezazu z_{beha} balio estatistiko bakoitzari dagokion p-balioa $H_0 : \mu = 30$ versus $H_1 : \mu \neq 30$ frogatzeko.

   a. 2.10
   b. –1.75
   c. -0.55
   d. 1.41
   e. –5.3

4. Ikertzaile batek $H_0 : \mu \leq 17$ versus $H_1 : \mu > 17$ frogatzeko datuak bildu ditu. Eskuartean dituzun banaketa-taulak erabiliz, borna ezazu p-balioa, emandako t_{beha} balio estatistikoa, eta dagokion askatasun-graduarentzat.

   a. $t_{beha} = 1.84$, askatasun-graduak = 14
   b. $t_{beha} = 3.74$, askatasun-graduak = 25
   c. $t_{beha} = 2.42$, askatasun-graduak = 13
   d. $t_{beha} = 1.32$, askatasun-graduak = 8
   e. $t_{beha} = 2.67$, askatasun-graduak = 45

5. Borna itzazu p-balioak bi aldeko t-proba batean 4. ariketaren antzera, honako kasu bakoitzean.

   a. $t_{beha} = 2.3$, askatasun-graduak = 6
   b. $t_{beha} = -3.0$, askatasun-graduak = 14
   c. $t_{beha} = 4.2$, askatasun-graduak = 24
   d. $t_{beha} = 1.3$, askatasun-graduak = 17

6. 11 lurretan fosforo kantitatea neurtzeko analisiak egin dira teknika berri bat erabiliz. Lur horietako bakoitzak duen benetako fosforo kantitatea 548 mg/kg-koa da. Laginaren batezbestekoa 587 mg/kg da, eta desbideratze estandarra 10.

   a. Fosforo kantitatea neurtzeko teknika berria egokia al da? Erabil ezazu $\alpha = 0.05$.
   b. Zer baldintza bete behar da aurreko ataleko froga zuzena izan dadin?

7. 16 alanbreko lagin bat hartu eta bakoitzari tentsio-erresistentzia (N/mm^2) froga bat eginda, laginaren batezbestekoa eta desbideratze estandarra 2160 eta 30 dira, hurrenez hurren.

   a. Beste teknika batez egindako alanbre batzuek 2150 N/mm^2-ko batez besteko tentsio-erresistentzia dute. Zein hipotesi frogatu beharko da gure alanbreen batez besteko tentsio-erresistentzia 2150 N/mm^2 baino handiagoa den argitzeko?

b. Tentsio-erresistentziaren banaketa normala dela suposatuz gero, zein estatistiko erabiliko zenuke aurreko ataleko hipotesia frogatzeko?
c. Zein da kasu hauetarako estatistikoaren behatutako balioa?
d. Ezar itzazu aurreko atalean kalkulatutako froga estatistikoaren balioari loturiko p-balioaren mugak.
e. Zer ondorioztatuko zenuke 0.05 mailako froga bat eginda?ç

8. Fabrikazioko aldakortasunak direla eta, altzairu lagin baten elastikotasun-muga erreala eta elastikotasun-muga teorikoa desberdinak dira. Demagun p elastikotasun-muga teorikora iritsi aurretik hausten diren piezen proportzio erreala dela. Elastikotasun-muga teorikora iritsi aurretik hausten den pieza kopurua % 20tik gorakoa dela onartuz gero, fabrikazio-prozesua aldatu egin beharko dela diote fabrikazio-arduradunek.

   a. Zorizko lagin bakun batean 60 piezatik 15 elastikotasun-muga teorikora iritsi aurretik hausten badira, zein da kasu honetan testaren p-balioa? Horren baitan, zein erabaki hartu behar lukete fabrikazio-arduradunek?
   b. Elastikotasun-muga teorikora iritsi aurretik hausten diren pieza-portzentaje erreala % 50ekoa izanik eta 0.01 mailako froga erabilita, zein da enpresak prozesua aldatu behar duela ondoriozta dezan probabilitatea?

9. Langargailu batzuekin sute-kasurako egindako froga batzuetan, langargailuek lehen aldiz martxan jartzeko behar izandako denbora honako hau da (segundotan):

27	41	22	27	23	35	30
33	24	27	28	22	24	

   Baldintza horietan sistemaren benetako aktibazio-denbora gehienez 25 segundokoa izan dadin diseinatua da sistema. Datuak sistemaren diseinu-espezifikazioarekin kontraesanean al daude?

   Idatz itzazu egoera horri dagozkion hipotesiak, eta, estatistiko egokiak erabiliz, lor itzazu lehengo datuen p-balioaren goi- eta behe-borneak. Zein erabaki hartuko duzu $\alpha = 0.05$ aukeratuz? Zer pentsatzen duzu aktibazio-denboraren banaketari buruz?

10. Karbono monoxido kontzentrazioa neurtzeko erabiltzen den espektrofotometro baten zehaztasuna gas industrial baten neurketa zehatzak eginda neurtzen da. Gas horren monoxido kontzentrazioa era oso zehatzean kontrolatzen da, eta 70 ppm-koa da espektrofotometroa ongi dabilenean. Neurketek espektrofotometroa ongi ez dabilela adierazten badute, berriz kalibratu beharko da. Kontzentrazio-neurketek banaketa normala dutela onartuz eta ondorengo neurketetan oinarriturik, beharrezkoa al da espektrofotometroaren berkalibrazioa? p-balioa kalkulatuz edo bornatuz eta $\alpha = 0.05$ aukeratuz, erantzun galderari.

Behatutako lagina: 85, 77, 82, 68, 72, 69.

11. Gobernuko funtzionarioak hondakin arriskutsuen ezabaketarekin lotutako arazoak ikertzen ari dira. Zabortegi baten inguruko 11 ur-putzutan laginak hartu dira. 38 eta 67 pCi/L arteko erradiazio-mailak aurkitu dira. Legeak edateko uretan gehienez 15 pCi/L onartzen du. Demagun 11 putzuetako laginak batez beste 52.5 pCi/L-ko erradiazioa dutela eta desbideratze estandarra 8 dela. Datu horien arabera eta 0.01 mailako froga erabiliz, esan daiteke batez besteko erradiazio-mailak legea hausten duela? Onar ezazu datuen normaltasuna.

12. Betaurrekoetan erabiltzen diren 50 lenteko lagin batean, 3.05 mm-ko batez besteko lodiera eta 0.34 mm-ko desbideratze estandarra neurtu dira. Lortu nahi den benetako batez besteko lodiera 3.20 mm-koa da. Lortutako informazioak adierazten al du lenteen egiazko batez besteko lodiera lortu nahi zenaren desberdina dela? Erabil ezazu $\alpha = 0.05$.

13. Lur batzuk ikatz-hautsarekin estali dira izotzaren aurkako tratamendu gisa. Horietako 8 lurretan fluxu termikoa neurtu da, eta honako balio hauek lortu dira:

1.5	5	-5.5	-2.7	-5.0	1.7	-1.1

Belardiek -1.7ko batez besteko fluxu termikoa dute. Fluxu termikoaren banaketa normala dela onartuz gero, eskuratutako datuek ikatz-hautsa fluxu termikoaren batezbestekoa handitzeko egokia dela adierazten al dute? Erabil ezazu $\alpha = 0.05$.

14. Ur-berogailuak egiten dituzten fabrika batean, bere berogailuek ura 37.8 $^{\circ}C$-ra berotzeko gehienez 15 minutu behar dituztela diote. Zoriz 32 berogailuko lagina hartu eta bakoitzak ura 37.8 $^{\circ}C$-ra berotzeko behar duen denbora neurtu da. Laginaren batez besteko denbora eta lagin desbideratze estandarra 17.5 minutu eta 2.2 minutu dira, hurrenez hurren. Datu horiek dudan jartzen al dute fabrikatik diotena? Kalkula ezazu p-balioa, 0.05 adierazgarritasun-mailako ondorio batera iristeko.

15. Populazioko banaketa normala denean eta n oso handia, lagin desbideratze estandarrak, gutxi gorabehera, banaketa normala dauka, $E(S) \approx \sigma$ eta $VAR(S) \approx \sigma^2/(2n)$ izanik. Badakigu, kasu honetan ere, $\bar{X}$ normala dela, $E(\bar{X}) = \mu$ eta $VAR(\bar{X}) = \sigma^2/n$ izanik.

    a. Oinarrizko banaketa normala dela suposatuz, zein litzateke 99. pertzentilaren ($\theta = \mu + 2.33\sigma$-ren) gutxi gorabeherako estimatzaile alboragabea?

    b. X_i-ak normalak direnean ($i = 1, \ldots, n$), $\bar{X}$ eta S zorizko aldagai askeak direla frogatu daiteke (batak kokapena neurtzen du eta besteak sakaba-

napena). Erabil ezazu hori $VAR(\hat{\theta})$ eta errore estandarra kalkulatzeko. Zein da estimatutako errore estandarra?

c. Idatz ezazu estatistiko bat $H_0 : \theta = \theta_0$ hipotesia frogatzeko, zeinak gutxi gorabeherako banaketa normal estandarra baitu H_0 egia denean.

Oharra: X eta Y zorizko aldagaien banaketa normala bada, $X + Y$ aldagaiaren banaketa ere normala da.

d. Lur bateko pH-ak banaketa normala izanik, 64 lur-laginetan $\bar{x} = 6.33$ eta $s = 0.16$ neurtu dira. Datu horiek argi adierazten al dute asko jota laginen % 99k pH=6.75 baino txikiagoa dutela? Hartu ezazu $\alpha = 0.01$ eta erabil ezazu atal honetan bertan planteatu duzun estatistikoa.

16. Gizonezko unibertsitarioak emakumezkoak baino errazago aspertzen al dira? Hori aztertzeko ikerketa bat egin zen, eta egileek Proneness aspertze-eskala bat aplikatu zieten 97 gizonezko eta 148 emakumezko ikasleri. Honako datu hauek gizonezkoetan eta emakumezkoetan *Proneness* aspertze-tasa berdina den hipotesiaren aldekoak al dira? Erabili hipotesi egokiak eta hartu erabakia 0.05eko adierazgarritasun-mailarekin.

	Lagin-tamaina	Batezbestekoa	Desbideratze estandarra
Gizonezkoak	97	10.40	4.83
Emakumezkoak	148	9.26	4.86

17. Ahotsaren errekonozimendurako bi sistema ditugu, A eta B. Bi sistemen errendimendua konparatzeko bi sistemak martxan jarri dira, eta honako datu hauek lortu:
    - A sistemarekin: n_1=160 hitz frogatu dira, eta 105 ongi identifikatu.
    - B sistemarekin: $n_2 = 192$ hitz frogatu dira, eta 135 ongi identifikatu.

    Informazio horrek, $\alpha = 0.05$-eko adierazgarritasun-mailaz, bi sistemen proportzioan diferentziarik adierazten al du?

A

Banaketa-taulak

Atal honetan, hainbat banaketa ezaguni dagozkien probabilitateak jaso dira, software estatistikoa eskura ez dugunerako.

Banaketa binomiala

Banaketa-funtzioak jaso dira: $X \sim Bin(n, p)$ izanik, $P(X \leq k)$ $(k = 0, \ldots, n)$.

		p										
n	**k**	**0.01**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**1/3**	**0.35**	**0.4**	**0.45**	**0.5**
2	**0**	0.980	0.902	0.810	0.723	0.640	0.563	0.444	0.423	0.360	0.302	0.250
	1	1.000	0.998	0.990	0.978	0.960	0.938	0.889	0.878	0.840	0.798	0.750
	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	**0**	0.970	0.857	0.729	0.614	0.512	0.422	0.296	0.275	0.216	0.166	0.125
	1	1.000	0.993	0.972	0.939	0.896	0.844	0.741	0.718	0.648	0.575	0.500
	2	1.000	1.000	0.999	0.997	0.992	0.984	0.963	0.957	0.936	0.909	0.875
	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	**0**	0.961	0.815	0.656	0.522	0.410	0.316	0.198	0.179	0.130	0.092	0.062
	1	0.999	0.986	0.948	0.890	0.819	0.738	0.593	0.563	0.475	0.391	0.313
	2	1.000	1.000	0.996	0.988	0.973	0.949	0.889	0.874	0.821	0.759	0.688
	3	1.000	1.000	1.000	0.999	0.998	0.996	0.988	0.985	0.974	0.959	0.938
	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

		p										
n	**k**	**0.01**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**1/3**	**0.35**	**0.4**	**0.45**	**0.5**
5	**0**	0.951	0.774	0.590	0.444	0.328	0.237	0.132	0.116	0.078	0.050	0.031
	1	0.999	0.977	0.919	0.835	0.737	0.633	0.461	0.428	0.337	0.256	0.187
	2	1.000	0.999	0.991	0.973	0.942	0.896	0.790	0.765	0.683	0.593	0.500
	3	1.000	1.000	1.000	0.998	0.993	0.984	0.955	0.946	0.913	0.869	0.812
	4	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.995	0.990	0.982	0.969
	5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6	**0**	0.941	0.735	0.531	0.377	0.262	0.178	0.088	0.075	0.047	0.028	0.016
	1	0.999	0.967	0.886	0.776	0.655	0.534	0.351	0.319	0.233	0.164	0.109
	2	1.000	0.998	0.984	0.953	0.901	0.831	0.680	0.647	0.544	0.442	0.344
	3	1.000	1.000	0.999	0.994	0.983	0.962	0.900	0.883	0.821	0.745	0.656
	4	1.000	1.000	1.000	1.000	0.998	0.995	0.982	0.978	0.959	0.931	0.891
	5	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.996	0.992	0.984
	6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7	**0**	0.932	0.698	0.478	0.321	0.210	0.133	0.059	0.049	0.028	0.015	0.008
	1	0.998	0.956	0.850	0.717	0.577	0.445	0.263	0.234	0.159	0.102	0.063
	2	1.000	0.996	0.974	0.926	0.852	0.756	0.571	0.532	0.420	0.316	0.227
	3	1.000	1.000	0.997	0.988	0.967	0.929	0.827	0.800	0.710	0.608	0.500
	4	1.000	1.000	1.000	0.999	0.995	0.987	0.955	0.944	0.904	0.847	0.773
	5	1.000	1.000	1.000	1.000	1.000	0.999	0.993	0.991	0.981	0.964	0.938
	6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.996	0.992
	7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8	**0**	0.923	0.663	0.430	0.272	0.168	0.100	0.039	0.032	0.017	0.008	0.004
	1	0.997	0.943	0.813	0.657	0.503	0.367	0.195	0.169	0.106	0.063	0.035
	2	1.000	0.994	0.962	0.895	0.797	0.679	0.468	0.428	0.315	0.220	0.145
	3	1.000	1.000	0.995	0.979	0.944	0.886	0.741	0.706	0.594	0.477	0.363
	4	1.000	1.000	1.000	0.997	0.990	0.973	0.912	0.894	0.826	0.740	0.637
	5	1.000	1.000	1.000	1.000	0.999	0.996	0.980	0.975	0.950	0.912	0.855
	6	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.996	0.991	0.982	0.965
	7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.996
	8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Poissonen banaketa

Banaketa-funtzioak jaso dira: $X \sim \mathcal{P}(\lambda)$ izanik, $P(X \leq k)$ $(k = 0, 1, 2, ...)$.

	λ													
k	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**1.1**	**1.2**	**1.3**	**1.4**
0	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407	0.368	0.333	0.301	0.273	0.247
1	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772	0.736	0.699	0.663	0.627	0.592
2	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937	0.920	0.900	0.879	0.857	0.833
3	1.000	1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987	0.981	0.974	0.966	0.957	0.946
4	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.996	0.995	0.992	0.989	0.986
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.998	0.997
6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

	λ													
k	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2**	**2.2**	**2.4**	**2.6**	**2.8**	**3**	**3.2**	**3.4**	**3.6**
0	0.223	0.202	0.183	0.165	0.150	0.135	0.111	0.091	0.074	0.061	0.050	0.041	0.033	0.027
1	0.558	0.525	0.493	0.463	0.434	0.406	0.355	0.308	0.267	0.231	0.199	0.171	0.147	0.126
2	0.809	0.783	0.757	0.731	0.704	0.677	0.623	0.570	0.518	0.469	0.423	0.380	0.340	0.303
3	0.934	0.921	0.907	0.891	0.875	0.857	0.819	0.779	0.736	0.692	0.647	0.603	0.558	0.515
4	0.981	0.976	0.970	0.964	0.956	0.947	0.928	0.904	0.877	0.848	0.815	0.781	0.744	0.706
5	0.996	0.994	0.992	0.990	0.987	0.983	0.975	0.964	0.951	0.935	0.916	0.895	0.871	0.844
6	0.999	0.999	0.998	0.997	0.997	0.995	0.993	0.988	0.983	0.976	0.966	0.955	0.942	0.927
7	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.995	0.992	0.988	0.983	0.977	0.969
8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.996	0.994	0.992	0.988
9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.996
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

	λ											
k	**3.8**	**4**	**4.2**	**4.4**	**4.6**	**4.8**	**5**	**5.2**	**5.4**	**5.6**	**5.8**	**6**
0	0.022	0.018	0.015	0.012	0.010	0.008	0.007	0.006	0.005	0.004	0.003	0.002
1	0.107	0.092	0.078	0.066	0.056	0.048	0.040	0.034	0.029	0.024	0.021	0.017
2	0.269	0.238	0.210	0.185	0.163	0.143	0.125	0.109	0.095	0.082	0.072	0.062
3	0.473	0.433	0.395	0.359	0.326	0.294	0.265	0.238	0.213	0.191	0.170	0.151
4	0.668	0.629	0.590	0.551	0.513	0.476	0.440	0.406	0.373	0.342	0.313	0.285
5	0.816	0.785	0.753	0.720	0.686	0.651	0.616	0.581	0.546	0.512	0.478	0.446
6	0.909	0.889	0.867	0.844	0.818	0.791	0.762	0.732	0.702	0.670	0.638	0.606
7	0.960	0.949	0.936	0.921	0.905	0.887	0.867	0.845	0.822	0.797	0.771	0.744
8	0.984	0.979	0.972	0.964	0.955	0.944	0.932	0.918	0.903	0.886	0.867	0.847
9	0.994	0.992	0.989	0.985	0.980	0.975	0.968	0.960	0.951	0.941	0.929	0.916
10	0.998	0.997	0.996	0.994	0.992	0.990	0.986	0.982	0.977	0.972	0.965	0.957
11	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.990	0.988	0.984	0.980
12	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.991
13	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Banaketa normala

Banaketa-funtzioak jaso dira: $Z \sim N(0,1)$ izanik, $P(Z \leq z)$, $z \geq 0$.

z	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Studenten t-banaketa

Banaketaren kuantilak (t_p) jaso dira taulan: k askatasun graduko $t \sim t_k$ banaketa izanik, $P(t \leq t_p) = p$.

k	0.995	0.99	0.975	0.95	0.9	0.8	0.75	0.7	0.6	0.55
1	63.657	31.821	12.706	6.314	3.078	1.376	1.000	0.727	0.325	0.158
2	9.925	6.965	4.303	2.920	1.886	1.061	0.816	0.617	0.289	0.142
3	5.841	4.541	3.182	2.353	1.638	0.978	0.765	0.584	0.277	0.137
4	4.604	3.747	2.776	2.132	1.533	0.941	0.741	0.569	0.271	0.134
5	4.032	3.365	2.571	2.015	1.476	0.920	0.727	0.559	0.267	0.132
6	3.707	3.143	2.447	1.943	1.440	0.906	0.718	0.553	0.265	0.131
7	3.499	2.998	2.365	1.895	1.415	0.896	0.711	0.549	0.263	0.130
8	3.355	2.896	2.306	1.860	1.397	0.889	0.706	0.546	0.262	0.130
9	3.250	2.821	2.262	1.833	1.383	0.883	0.703	0.543	0.261	0.129
10	3.169	2.764	2.228	1.812	1.372	0.879	0.700	0.542	0.260	0.129
11	3.106	2.718	2.201	1.796	1.363	0.876	0.697	0.540	0.260	0.129
12	3.055	2.681	2.179	1.782	1.356	0.873	0.695	0.539	0.259	0.128
13	3.012	2.650	2.160	1.771	1.350	0.870	0.694	0.538	0.259	0.128
14	2.977	2.624	2.145	1.761	1.345	0.868	0.692	0.537	0.258	0.128
15	2.947	2.602	2.131	1.753	1.341	0.866	0.691	0.536	0.258	0.128
16	2.921	2.583	2.120	1.746	1.337	0.865	0.690	0.535	0.258	0.128
17	2.898	2.567	2.110	1.740	1.333	0.863	0.689	0.534	0.257	0.128
18	2.878	2.552	2.101	1.734	1.330	0.862	0.688	0.534	0.257	0.127
19	2.861	2.539	2.093	1.729	1.328	0.861	0.688	0.533	0.257	0.127
20	2.845	2.528	2.086	1.725	1.325	0.860	0.687	0.533	0.257	0.127
21	2.831	2.518	2.080	1.721	1.323	0.859	0.686	0.532	0.257	0.127
22	2.819	2.508	2.074	1.717	1.321	0.858	0.686	0.532	0.256	0.127
23	2.807	2.500	2.069	1.714	1.319	0.858	0.685	0.532	0.256	0.127
24	2.797	2.492	2.064	1.711	1.318	0.857	0.685	0.531	0.256	0.127
25	2.787	2.485	2.060	1.708	1.316	0.856	0.684	0.531	0.256	0.127
26	2.779	2.479	2.056	1.706	1.315	0.856	0.684	0.531	0.256	0.127
27	2.771	2.473	2.052	1.703	1.314	0.855	0.684	0.531	0.256	0.127
28	2.763	2.467	2.048	1.701	1.313	0.855	0.683	0.530	0.256	0.127
29	2.756	2.462	2.045	1.699	1.311	0.854	0.683	0.530	0.256	0.127
30	2.750	2.457	2.042	1.697	1.310	0.854	0.683	0.530	0.256	0.127
40	2.704	2.423	2.021	1.684	1.303	0.851	0.681	0.529	0.255	0.126
60	2.660	2.390	2.000	1.671	1.296	0.848	0.679	0.527	0.254	0.126
120	2.617	2.358	1.980	1.658	1.289	0.845	0.677	0.526	0.254	0.126

Khi-karratuaren banaketa

Banaketaren kuantilak (q_p) jaso dira taulan: k askatasun graduko $\chi^2 \sim \chi^2_k$ banaketa izanik, $P(\chi^2 \leq q_p) = p$.

k	0.995	0.99	0.975	0.95	0.9	0.8	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	7.9	6.6	5.0	3.8	2.7	1.6	1.3	0.5	0.1	0.0	0.0	0.0	0.0	0.0
2	10.6	9.2	7.4	6.0	4.6	3.2	2.8	1.4	0.6	0.2	0.1	0.1	0.0	0.0
3	12.8	11.3	9.3	7.8	6.3	4.6	4.1	2.4	1.2	0.6	0.4	0.2	0.1	0.1
4	14.9	13.3	11.1	9.5	7.8	6.0	5.4	3.4	1.9	1.1	0.7	0.5	0.3	0.2
5	16.7	15.1	12.8	11.1	9.2	7.3	6.6	4.4	2.7	1.6	1.1	0.8	0.6	0.4
6	18.5	16.8	14.4	12.6	10.6	8.6	7.8	5.3	3.5	2.2	1.6	1.2	0.9	0.7
7	20.3	18.5	16.0	14.1	12.0	9.8	9.0	6.3	4.3	2.8	2.2	1.7	1.2	1.0
8	22.0	20.1	17.5	15.5	13.4	11.0	10.2	7.3	5.1	3.5	2.7	2.2	1.6	1.3
9	23.6	21.7	19.0	16.9	14.7	12.2	11.4	8.3	5.9	4.2	3.3	2.7	2.1	1.7
10	25.2	23.2	20.5	18.3	16.0	13.4	12.5	9.3	6.7	4.9	3.9	3.2	2.6	2.2
11	26.8	24.7	21.9	19.7	17.3	14.6	13.7	10.3	7.6	5.6	4.6	3.8	3.1	2.6
12	28.3	26.2	23.3	21.0	18.5	15.8	14.8	11.3	8.4	6.3	5.2	4.4	3.6	3.1
13	29.8	27.7	24.7	22.4	19.8	17.0	16.0	12.3	9.3	7.0	5.9	5.0	4.1	3.6
14	31.3	29.1	26.1	23.7	21.1	18.2	17.1	13.3	10.2	7.8	6.6	5.6	4.7	4.1
15	32.8	30.6	27.5	25.0	22.3	19.3	18.2	14.3	11.0	8.5	7.3	6.3	5.2	4.6
16	34.3	32.0	28.8	26.3	23.5	20.5	19.4	15.3	11.9	9.3	8.0	6.9	5.8	5.1
17	35.7	33.4	30.2	27.6	24.8	21.6	20.5	16.3	12.8	10.1	8.7	7.6	6.4	5.7
18	37.2	34.8	31.5	28.9	26.0	22.8	21.6	17.3	13.7	10.9	9.4	8.2	7.0	6.3
19	38.6	36.2	32.9	30.1	27.2	23.9	22.7	18.3	14.6	11.7	10.1	8.9	7.6	6.8
20	40.0	37.6	34.2	31.4	28.4	25.0	23.8	19.3	15.5	12.4	10.9	9.6	8.3	7.4
21	41.4	38.9	35.5	32.7	29.6	26.2	24.9	20.3	16.3	13.2	11.6	10.3	8.9	8.0
22	42.8	40.3	36.8	33.9	30.8	27.3	26.0	21.3	17.2	14.0	12.3	11.0	9.5	8.6
23	44.2	41.6	38.1	35.2	32.0	28.4	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.3
24	45.6	43.0	39.4	36.4	33.2	29.6	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.9
25	46.9	44.3	40.6	37.7	34.4	30.7	29.3	24.3	19.9	16.5	14.6	13.1	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	31.8	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	32.9	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	34.0	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.6	39.1	35.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	36.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	47.3	45.6	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	58.2	56.3	49.3	42.9	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	69.0	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	85.5	79.7	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	90.4	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	101.1	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	111.7	109.1	99.3	90.1	82.4	77.9	74.2	70.1	67.3

Bibliografia

[1] Alan Agresti eta Christine A Franklin. *Statistics: The art and science of learning from data*. Pearson, 2. argit., 2009.

[2] Inmaculada Arostegui eta Arantza Urkaregi. *Bioestatistika: Oinarrizko Ikastaroa*. Euskal Herriko Unibertsitatea, Argitalpen Zerbitzua, 2014.

[3] Michael Baron. *Probability and Statistics for Computer Scientists*. CRC Press, 3. argit., 2019.

[4] Peter Dalgaard. *Introductory Statistics with R*. Springer, 2008.

[5] Jay L. Devore. *Probabilidad y Estadística para Ingeniería y Ciencias*. International Thomson, 2001.

[6] J. Susan Milton eta Jesse C. Arnold. *Probabilidad y Estadística con Aplicaciones para Ingeniería y Ciencias Computacionales*. McGraw-Hill, 4. argit., 2004.

[7] Douglas C. Montgomery eta George C. Runger. *Probabilidad y Estadística Aplicadas a la Ingeniería*. Limusa Wiley, 2. argit., 2004.

[8] William Navidi. *Estadística para Ingenieros*. McGraw-Hill / Interamericana, 2006.

[9] Daniel Peña. *Fundamentos de Estadística*. Alianza Editorial, 2010.

[10] Vicente Quesada, Angel Isidoro, eta Luis López. *Curso y Ejercicios de Estadística*. Alhambra Universidad, 1982.

[11] Venancio Tomeo eta Isaías Uña. *Lecciones de Estadística Descriptiva*. Thomson, 2003.

[12] Kishor S. Trivedi. *Probability ad Statistics with Reliability, Queuing, and Computer Science Applications*. Wiley, 2001.

[13] Isaías Uña, Venancio Tomeo, eta Jesús San Martín. *Lecciones de Cálculo de Probabilidades*. Thomson, 2003.

[14] Ronald E Walpole, Raymond H Myers, eta Sharon L Myers. *Probabilidad y estadística para ingenieros*. Pearson educación, 1999.