

ESTATISTIKA ETA DATUEN ANALISIA

I. ikasgaia: Estatistikaren hastapenak

- 1.1 Estatistika zer den eta zer ez den
- 1.2 Estatistika eta zientziaren paradigma positibista
- 1.3 Estatistika eta ikerketa nomotetikoa
- 1.4 Estatistika eta metodo induktiboa
- 1.5 Errore estatistikoa: zorizko errorea eta lagin-errorea
- 1.6 Estatistika deskribatzailea eta inferentzia estatistikoa
- 1.7 Estatistikaren historiari gainbegirada bat
- 1.8 Estatistikaren aplikazio arloak
- 1.9 Ikerketa estatistikoaren plangintza
- 1.10 Aldagai estatistikoak

Egilea: Josemari Sarasola



Gizapedia

gizapedia.org

1. gaia: Estatistikaren hastapenak

1.1 Estatistika zer den eta zer ez den

Hiztegian *estatistika* hitza bilatzen dugunean, bi adiera aurkitzen ditugu:

- lehen adiera batean, estatistika datu-multzo bat da, eta hala, "langabeziaren estatistikak" edota "hauteskundeetako estatistikak" aipatzen dira, besteak beste;
- bigarrenean, estatistika datu-multzoak kuantitatiboki aztertzen dituen diziplina edo jakintza-arloa da.

Bigarren adiera da interesatzen zaiguna, sakondu behar duguna, diziplina horren *objektu materiala* eta *objektu formala* bereiziz. Estatistikaren objektu materiala datuak eta datu multzoak dira, estatistikaren aztergaia alegia, eta objektu formala datu horiekin, zer eta nola egin behar dugun. Horrela, esan daiteke **estatistika dela errealitateko fenomeno aldakor bati buruzko datu multzoak (objektu materiala) modu egokian jaso eta aztertzen dituzten tekniken multzoa, horretarako datuetan dauden egiturak, joerak eta erregulartasunak, ondorio orokor gisa, bilatuz (objektu formala).**

Definizio sinple eta zehatz horretan hainbat alderdi nabarmendu behar dira. Lehenik eta behin, estatistikaren **objektu materiala** edo aztergaia aipatu behar da: *fenomeno aldakorrak*, elementu edo aldi bakoitzean balio ezberdinak hartzen dituzten datu-multzotan irudikatuak; adibidez, estatistikaren aztergaiak lirateke ikasleen kalifikazioak, familien errentak eta egunez eguneko tenperaturak; baina inondik ere ez, balio finkoa hartzen dutenez, autobus baten irteera ordua, ikasleek irakasgaia gainditzeko behar duten nota eta diru-laguntza jasotzeko behar den errenta minimoa.

Objektu formalari dagokionean, estatistikak datu-multzoei buruzko ondorio orokorrak ematea da, eredu, sintesi edo laburpenak alegia. Kontuz interpretatu

behar dira estatistikak ematen dituen ondorio horiek; adibidez, baieztatzen dugunean ikasleen batez besteko kalifikazioa 7 izan dela, horrek ez du esan nahi ikasle guztiek gainditu dutenik; 7 puntuko kalifikazioa balio zentrala bat dela baizik, joera bat alegia, horren gainetik eta azpitik beti egongo baitira ikasleak. Halaber, ez ditu datu bakanak edo kasu partikularrak aztertzen edota haiei buruz konklusioak ateratzen, haietarako aurreanak egiteko lagungarria bada ere (gogoratu euskal esaera zaharra: "enara batek ez du udaberria egiten").

Beste alde batetik, estatistikaren emaitzak jasotzen ditugunean, argi geratu behar da estatistika ez dela erabili behar kalkulagailu huts gisa. Estatistikaren kalkulagailuak idatzen dituen emaitzak irakurtzen jakin behar da. Emaitzak beti interpretatu behar dira.

Azkenik, lan estatistikoa zuzen burutzeko datuak behar bezala jaso behar dira. Ildo horretan, nabarmendu behar da *behar diren datuak soilik* jaso behar direla, ez gehiago eta ez gutxiago, eta horretarako lan estatistikoaren helburua aurrez zehaztea komeni da. Datu gutxiegia jasotzen badira, datu-multzoaren azterketatik jasoko den informazioa eskasa izango da. Baina datu gehiegia jasotzea ere kaltegarri izaten da, lehenik eta behin arrazoi ekonomikoengatik, datuak jasotzeak kostu bat dakarrelako, baina bereziki datu gehiegia jasotzeak fenomeno argitu bainoago datu-masa handian aztergai dugun fenomeno nahasi eta ilun azaltzen zaigulako. Datu kopuru handiegiaren desegokitasuna ilustratzeko, Jorge Luis Borgesek ipuin labur bat, *Zientziaren zehaztasunaz* ("Del rigor en la ciencia") izenburukoa, dakargu gogora:

Inperio hartan, Kartografiaren Arteak hainbestearainoko Perfekziora heldu zen, non Probintzia bakar bateko Mapak Hiri oso baten lekua hartzen baitzuen, eta Probintzia oso bat Inperioaren Mapak. Denboraren poderioz, Neurrigabeko Mapa horiek ez ziren nahiko, eta Kartografoen Elkargoek Inperioaren Tamaina zuen Inperioaren Mapa eratu zuten, harekin guztiz bat zetorrena. Kartografiaren Ikasketarekin hain zaletuak ez zirenez, Ondorengo Belaunaldiek Mapa zabal hura Baliogabea zela jakin zuten, eta Eguzkiaren eta Neguaren gorabeheren mendean utzi zuten. Mendebaldeko Basamortuetan Maparen Hondakinek puskatuta diraute, Pizti eta Arloteen bizileku; Herrialde osoan ez dago Diziplina Geografikoen arrasto gehiagorik.

Narrazio labur honek datu gehiegia hartzearen eragozpenez adierazteaz gainera (kasu horretan, neurrigabeko mapa bat egitea), argi uzten ditu estatistikaren helburua: **fenomeno aldakorrek sinplifikatzea, haietan dauden joera, erregulartasun eta beste informazio jakingarriak jasotzea**. Izan ere, Borgesek Neurrigabeko Mapak bezala, datu-multzo gordina, hortik behar den informazioa atera gabe, errealitatearen kopia hutsa da, inongo baliorik ez duena; azkenean, errealitatea ezagutzeko, hura laburtu egin behar baita.

1.2 Estatistika eta zientziaren paradigma positibista

Gizarte zientzietan, hainbat paradigma edo ikuskera daude ikerketa egiteko, hala nola paradigma sozio-kritikoa, paradigma fenomenologikoa, paradigma estrukturalista eta abar. Paradigma positibista natur zientzietako paradigma nagusia da, eta giza zientzietara ere zabaldu da, gizarte zientzietan ere natur zientzietako ezagutza-maila lortze aldera. Paradigma horren funtsa da ikerketa enpirikoa da, neurketaz eta behaketaz jasotako datuen azterketa, eta alde horretatik esan daiteke estatistika dela paradigma horretako tresna nagusietako bat.

1.3 Estatistika eta ikerketa nomotetikoak

Fenomeno bat aztertzen dugunean, bi ikerketa-modu ditugu aukeran: **ikerketa idiografikoa** eta **ikerketa nomotetikoa**. Kasu partikular baten azterketa egin dezakegu, kasu hori berez interesgarria delakoan, edota bere ondorioak fenomeno orokorrago batera zabal daitezkeela uste dugulakoan; adibidez, Donostiako etxebizitzaren merkatua 2022 urtean aztertzen dugunean. Kasu partikular bakanak aztertzen dituzten horiei ikerketa idiografiko deritze.

Ikerketa nomotetikoan berriz, fenomeno bat bere aniztasunean eta aldakortasunean aztertzen dugu, fenomenoak osatzen duten kasu guztiak aztertuz, eta kasu guztien azterketatik ondorio orokorrak eskuratuz. Argi dago estatistikaren zerregina ikerketa nomotetiko horietan kokatzen dela, fenomenoaren kasu guztiak batera aztertzeke teknika gisa.

1.4 Estatistika eta metodo induktiboa

Zientzia garatzeko bi metodo ditugu: *dedukzioa*, axioma edo baieztapen batzuetatik beste baieztapen batzuk logikaz ondorioztatzen dituen (matematikaren erabiltzen da bereziki); eta *indukzioa*, kasu indibidual edo partikularretatik lege orokorrak eratzen dituen, gehienetan aldakortasuna onartuz. Adibidez, herrialde guztietako datuak bilduz, industriaren garrantzia zenbat eta handiagoa izan, *orokorrean* langabezia orduan eta txikiagoa dela ikisten badugu, industriak langabezia gutxitzeko sektore garrantzitsuena dela esan daiteke. Arestiko baieztapena kasu partikularrak aztertuz egin da, eta horregatik esaten da horretan metodo induktiboa garatu dugula. Indukzioak darabilen metodoa estatistika da: estatistika da, hain zuzen, banako datuak aztertu eta haietatik fenomeno edo populazio oso bati buruzko ondorio orokorrak ematen dituen.

1.5 Errore estatistikoa: zorizko errorea eta lagin-errorea

Estatistikak fenomeno aldakorak aztertu eta konklusio zehatzak ematen dituzenez, aurrean bat egitean errore bat sortzen da estatistikak baieztatzen duenaren eta benetan gertatzen denaren artean. Adibidez, ikasle baten batez besteko nota 6 dela adierazten denean, eta emaitza hori azterketa egin behar duen ikasle baten nota aurreateko erabiltzen denean, errore bat sortuko da ziur aski.

Errore estatistikoak bi sorburu ditu:

- **errore estokastikoa edo zorizko errorea**, datu estatistikoak berez aldatkorak direlako. Aldakortasun horrek bi jatorri izan ditzake: fenomenoak eragina duten *faktore ezezagun eta kontrolagaitzak*, alde batetik; eta fenomenoaren berezko *zorizkotasuna*. Adibidez, ikaslearen nota 6 izango dela aurreatean, batezbestekoan soilik oinarrituta, ikaslearen notan eragina duten beste faktore batzuk baztertzeko dira (zenbat ordu ikasi dituen, maila sozioekonomikoa, eta abar), haiek kontuan hartuz gero aurrean zehatzagoa emango luketenak; horretaz gainera, eta zorizkotasunari buruz, ikasle baten azterketaren notan eragina duten faktore guztiak ikertuta ere, ikasleak zorte ona edo txarra izan dezake azterketan, nota handiagoa edo txikiagoa, aurreana egitean errorea ekarriko duena;
- **lagin-errorea**; izan ere, aztergai den populazio osoaren ordean, populazioari buruzko ondorioak ateratzeko haren azpimultzo bat, *lagin* izenekoa, aztertzeko da maiz. Lagina populazioaren adierazgarri izan dadin, lagineko elementuak zoriz jaso behar dira (adibidez, azokan intxaurren kalitateari buruz jakiteko, ez ditugu soilik gaineko intxaurreak aztertuko da, multzo osotik aukeratu behar dira, eta horretarako bide egokiena intxaurreak zoriz aukeratzeko da). Emaitzak laginetik populaziora zabaltzean, errore bat sortzen da, zeren lagina, zoriz jasota ere, ez baitu erabateko zehaztasunez populazioa islatzen. Horregatik, komeni da datu-multzoa populazioaren lagina denean, eta handik konklusioak ateratzen direnean *lagin-errorearen erreserbapean* esamoldea gehitzea haiei. Azkenik, ohartu behar da lagina zenbat eta handiagoa den, lagin-errorea orduan eta txikiagoa izango dela.

1.6 Estatistika deskribatzailea eta inferentzia estatistikoa

Datu-multzoak besterik gabe deskribatu egin nahi direnean, haiek irudikatzen dituzten grafikoak eratuz edota kalkulu sinpleak (batezbestekoak, esate baterako) eginez, errore estatistikoa kontuan hartu gabe, *estatistika deskribatzailea* egiten da.

Aldiz, datu-multzoetatik ateratako ondorioetan dagoen errore estatistikoa zenbatetsi eta kontrolatzeko teknikak badaude, *probabilitate-teorian* oinarritzen direnak. Errore estatistikoa aztertu egiten duen estatistikaren adarrari *inferentzia estatistikoa* deritzo.

1.7 Estatistikaren historiari gainbegirada bat

Estatistika datu-bilduma huts bezala ulertzen bada, antzinatek praktikatzen da. Antzinaroko zibilizazioetan (Txinan, Mesopotamian, Egipton eta Erroman, kasu) ohikoak ziren zentsuak eta antzeko datu-bilketak; inken eta Andeetako beste kulturen artean, adibidez, *quipu* izeneko soka-sorta datuak lokarrien bitartez erregistratzeko tresna gisa erabiltzen zen. Analisisirako tresna gisa, berriz, estatistikaren sorrera XVII. mendearen bigarren erdialdean kokatzen da, John Graunt-en eskutik, Londreseko hilkortasun-tasak aztertuz hiri hartako biztanleriaren zenbatespen edo estimazio bat egiteko metodo bat asmatu zuenean.

XIX. mendera arte, ordea, estatistika datu-bilketa geografiko, politiko eta ekonomikoen bilduma huts bezala ulertu zen orokorrean. Erresuma Batuan bereziki, *aritmética política* deitu zitzaion herrialdeak datuen bidez deskribatzeko modu horri. 1749 urtean, ordea, Gottfried Achenwall ekonomialari alemaniarrek *Statistik* terminoa asmatu zuen, italierazko *statista* edo politikari hitzetik, herrialdeen datu bilketa deskribatzaile horiek izendatzeko, eta termino hori izan zen azkenean nagusitu zena.

XIX. mendean, estatistikaren eremua arlo geografiko eta ekonomikotik beste arlo batzuetara zabaldu zuen, gizarte zein natur zientzietara. Hain zuzen, Adolphe Quetelet-ek *batez besteko gizakia*-ren ideia zabaldu zuen XIX. mendearen erdialdean, gizartearen errealitate konplexua ulertze aldera.

Aldi berean, XVII. mendetik probabilitatearen teoria garatzen joan zen, gehienetan zorizko jokoak aztertzeke, baina datuak aztertzeke tresna gisa izan zena.



Irudia 1.1: Quipu, soketan lokarriak eginez, datu-bilketarako tresna.

keen balioaz oihartu gabe. XIX. mendean, ordea, estatistika probabilitatearen kontzeptua barneratzen joan zen, datuen aldakortasuna, zorizko errorea eta lagin errorea azaltzearen, eta hala probabilitatearen kalkulua barneratuz, metodo estatistikoek zorrotasun matematikoa eskuratzen joan ziren. Bide horretatik XX. mendearen erdialderako *estatistika matematikoa*, probabilitate-teoriarekin bat egiten duen estatistika alegia, ia guztiz garatuta geratu zen.

XX. mendearen bigarren erdialdean informatikaren garapena gertatu zen, estatistikari bide berriak zabaltu zizkiona, bereziki datu-multzo handiak aztertzeke. Garai horretan garatzen da *aldagai anitzeko analisia*, aldagai askotako datuak batera aztertzeke metodoak garatzen dituenak. Interneten nagusitzearekin batera, datu multzo itzelak sortzen dira, modu egokian bildu eta prozesatu behar direnak (*big data* deitzen zaio arlo honi). Datu-multzo itzel horietatik informazio jakingarria eskuratzeko metodologiari *data mining* edo *datu-meatzaritza* deitzen zaio.

1.8 Estatistikaren aplikazio-arloak

Jakintzaren arlo guztietan aplikatu daiteke estatistika, astronomiatik (izar mota ezberdinen ezaugarriak aztertzeke) literatura (idazle ezberdinen estiloak era kuantitatiboan alderatzeko); esan liteke estatistika *mathesis universalis* dela, zientzia unibertsal bat alegia, unibertsoko eta munduko fenomeno guztiak ikerzteke balio duena. Horren frogarria estatistika unibertsitateko gradu gehienetako ikasketen irakasgai moduan agertzea da. Dena den, aplikatzen den arloa zein den, metodo estatistiko bereziak erabiltzen dira. Hala, jakintza-arlo batzuetako metodo estatistikoek izen berezia hartu dute:

- **ekonometria**, ekonomiara aplikaturiko estatistika da;
- **bioestatistika**, biologian eta medikuntzan garatzen dena;
- **epidemiologia**, gaixotasunen maiztasuna ikertzen duena;
- **psikometria**, psikologiara aplikaturiko estatistika (testak nola eratu behar diren aztertzen du, besteak beste);
- **demografia**, populazioak eta horien bilakaera aztertzen dituenak;
- **astrometria**, astronomiara aplikaturiko estatistika;
- **geoestatistika**, meteorologia, klimatologia, geologia eta beste luraren zientzietara aplikatzen dena eta bereziki denborazko eta espaziozko datuak aztertzen dituenak.

1.9 Ikerketa estatistikoaren plangintza

Datuak modu egokian jasotzen hasi aurretik, garrantzitsua ikerketaren hainbat alderdi zehaztea, zein datu bildu eta nola jaso behar diren erabakitzeke: Ikerketaren helburua zehaztu behar da aurretik, *zer jakin nahi dugun* alegia. Helburua zein den, halako teknika estatistikoa baliatu beharko da. Estatistikak eman ditzakeen konklusioak era askotakoak dira, baina maiz honako multzo hauetako batean biltzen da:

- *konparaketak* egitea datu-multzo ezberdinak hartuta; adibidez, lantegi batean diharduten emakumeek eta gizonek oro har ekoizpen ezberdina izaten duten eta zenbateraino;
- *erlazioak* bilatzea aldagaien artean; adibidez, publizitate-gastuak salmentak handitzen dituen ala ez, eta zenbateraino;
- *aurresanak* egitea, iraganeko denbora-serie batean oinarrituz etorkizuneko denbora batean izango den balio bati buruz zenbatespen bat ematea alegia.

Ondoren, *ikerketa-unitatea* (zein objektu edo entitate mota ikertuko den); *behaketa-unitatea* (datuak zeren gainean behatu edo jasoko diren), *aldagaia*, zer jaso edo neurtuko dugun, *espazio-eremua*, nongo datuak jasoko diren, eta *denbora-eremua*, noizko datuak bilduko diren, zehatz definitu behar dira, jaso behar ditugun datuak zein diren zehaztasunez jakiteko. Adibidez, 2015eko Gipuzkoako familien errenta aztertu nahi bada, *familia* zehazki definitu beharko da, ikerketa-unitate gisa; *behaketa-unitatea* berriz, pertsona fisikoa izango da, familia bakoitzaren errenta kalkulatzeko, familia horretako kide bakoitzaren errenta jaso behar delako (askotan behaketa-unitatea eta ikerketa-unitatea gauza bera dira, baina ezberdinak ere izan daitezke, adibide honetan kasu); orobat zehaztu beharko da zer esan nahi dugun *Gipuzkoako* esaten dugunean (Gipuzkoako erroldan egotea, Gipuzkoan lan egitea,...), espazio-eremua alegia; eta baita ere *2015eko errenta* jasotzean zer adierazi nahi den zehazki, aldagaia (errenta) eta denbora-eremua (2015) alegia. Kontzeptu horiek guztiak behar bezala argitzen ez badira, datu *heterogeneoak* eskuratzeko arriskua dago, gauza berari buruzkoak ez izatekoa alegia. Denborari dagokionean, azkenik, ohartu behar da hurrengo ikasgaietan serie edo datu-multzo estatikoak aztertu ditugula, une edo epe berean jasotakoak, eta hala ez denean, denbora ez dela faktore garrantzitsua izango.

1.10 Aldagai estatistikoak

Aldagaia datuetan jaso eta neurtzen dena da. Adibidez, Gipuzkoako familien errentari buruzko datuak biltzean, aldagaia errenta da. Nolako aldagaiak ditugun, halako metodo estatistikoak garatu beharko dira. Horregatik da garrantzitsua jakitea nolako aldagaia dugun aurrean:

- **Aldagai kualitatibo edo kategorikoak, atributu** ere deituak, kualitate bat (eta ez kopuru edo neurri bat) jasotzen dutenak; adibidez, ikasle baten kalifikazioa (eskas, nahiko, ongi, oso ongi, bikain) eta sexua (neska edo mutil). Beraz, aldagai kualitatiboek zer? edo nolakoa? galderei erantzuten diete. Horietan beste bereizketa bat egin daiteke:
 - **Aldagai nominalak**, ordena eman ezin daitekeenean (ikasle batek egiten duen gradua, adibidez); horietan, **aldagai dikotomikoak** bereizten dira, bi modalitate edo kategoria bakarrik jasotzen dituztenak (adibidez, sexua: gizon/emakume).
 - **Aldagai ordinalak**, kategoria ezberdinak mailakatzen direnean; adibidez, ikaslearen kalifikazioa (eskas, nahiko, ...). Ohikoa da, halaber, aldagai ordinalak elementu batzuen arteko preferentziak adierazteko (lanpostu baterako hautagai batzuk ordenatzen direnean, adibidez).
- **Aldagai kuantitatiboak**, zenbaki batez *neurtu* egiten dutenean (adibidez, adina urtetan). Hau da, aldagai kuantitatiboak zenbat galderari erantzuten dio.

Nabarmendu hartu behar da aldagaiak ez dira berez kuantitatiboak edo kualitatiboak: nola jaso edo neurtzen diren, halakoak izango dira. Adibidez, matematika-nota kuantitatiboa (8.7) zein kualitatiboa (oso ongi) izan daiteke.