ESTATISTIKA ETA DATUEN ANALISIA

I. ikasgaia: Estatistikaren hastapenak

Egilea: Josemari Sarasola

Gizapedia

gizapedia.hirusta.io

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 aztertzen dituen jakintza-arloa da.

Bigarren adiera da interesatzen zaiguna, eta horregatik zehaztu beharrean gaude horren definizioa, datuekin zer eta nola egin behar dugun jakiteko. Honela adieraz daiteke **estatistika zer den: errealitateko fenomeno aldakor bati buruzko datu multzoak modu egokian jaso eta aztertzen dituzten tekniken multzoa, horretarako datuetan dauden joerak eta erregulartasunak bilatuz**.

Definizio sinple eta zehatz horretan hainbat alderdi nabarmendu behar dira. Lehenik eta behin, estatistikaren aztergaia aipatu behar da: *fenomeno aldakorrak*, elementu edo aldi bakoitzean balio ezberdinak hartzen dituzten datumultzotan adieraziak; 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.

Bigarrenik, lan estatistikoa zuzen burutzeko datuak behar bezala jaso behar direla adierazten da definizioan. 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 gutxiegi jasotzen badira, datu-multzoaren azterketatik jasoko den informazioa eskasa izango da. Baina datu gehiegi jasotzea ere kaltegarri izaten da, lehenik eta behin arrazoi ekonomikoengatik, datuak jasotzeak kostu bat dakarrelako, baina bereziki datu gehiegi jasotzeak fenomenoa argitu bainoago datumasa handian aztergai dugun fenomenoa nahasi eta ilun azaltzen zaigulako. Datu kopuru handiegiaren desegokitasuna ilustratzeko, Jorge Luis Borgesen ipuin labur bat, *Zientziaren zehaztasunaz* ("Del rigor en la ciencia") izenburukoa, ekar dezakegu gogora:

> Inperio hartan, Kartografiaren Artea hainbesterainoko 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 gehiegi hartzearen eragozpenez adierazteaz gainera (kasu horretan, neurrigabeko mapa bat egitea), argi uzten ditu estatistikaren helburua: **fenomeno aldakorrak sinplifikatzea, haietan dauden joera, erregulartasun eta beste informazio jakingarria jasotzea**. *Konklusioak* eman beharra, alegia. Izan ere, Borgesen Neurrigabeko Mapak bezala, datumultzo 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, metodo induktiboaren tresna

Egun, ezinbestekotzat hartzen da estatistikaren erabilera teoria eta hipotesi zientifikoak baieztatu eta frogatzeko. Zientziaren metodoak bi dira funtsean: *dedukzioa*, axioma edo baieztapen batzuetatik beste baieztapen batzuk logikaz ondorioztatzen dituena (matematikan erabiltzen da bereziki); eta *indukzioa*, kasu indibidual edo partikularretatik lege orokorrak eratzen dituena, gehienetan aldakortasuna onartuz. Indukzioak darabilen metodoa estatistika da: estatistika da, hain zuzen, banako datuak aztertu eta haietatik fenomeno edo populazio oso bati buruzko ondorioak ateratzen dituena.

1.3 Errore estatistikoa: behaketa-errorea eta lagin-errorea

Estatistikak fenomeno aldakorrak aztertu eta konklusio finkoak ematen dituenez, aurresan 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 aurresateko erabiltzen denean, errore bat sortuko da ziur aski.

Errore estatistikoak bi sorburu ditu:

- **errore estokastikoa**, datu estatistikoak berez aldakorrak 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 aurresatean, batezbestekoan soilik oinarrituta, ikaslearen notan eragina duten beste faktore batzuk baztertzen dira (zenbat ordu ikasi dituen, maila sozioekonomikoa, eta abar), haiek kontuan hartuz gero aurresan 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, aldakortasuna alegia, ekarriko diona;

- **lagin-errorea**; izan ere, aztergai den populazio osoaren ordez, populazioari buruzko ondorioak ateratzeko haren azpimultzo bat, *lagin* izenekoa, aztertzen da maiz. Lagina populazioaren adierazgarri izan dadin, lagineko elementuak zoriz jaso behar dira (adibidez, azokan intxaurren kalitateari buruz jakiteko, ez ditugu soilik gaineko intxaurrak aztertuko da, multzo osotik aukeratuko dira, eta horretarako bide egokiena intxaurrak zoriz aukeratzea 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 lagin-errorea orduan eta txikiagoa izango dela, lagina zenbat eta handiagoa den.

1.4 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.

1.5 Estatistikaren historiari gainbegirada bat

Estatistika datu-bilduma huts bezala ulertzen bada, antzinatik praktikatzen da. Antzinako zibilizazioetan (Antzinako Txinan, Antzinako Egipton eta Antzinako Erroman, kasu) ohikoak ziren zentsuak eta antzeko datu-bilketak. Analisirako tresnatzat harturik, berriz, estatistikaren sorrera XVII. mendearen bigarren erdialdean kokatu behar da John Graunten eskutik, Londreseko hilkortasun-tasak aztertuz hiri hartako biztanleriaren zenbatespen edo estimazio bat egiteko metodo bat plazaratu zuena. XIX. mendera arte, ordea, estatistika datu-bilketa geografiko, politiko eta ekonomikoen bilduma huts bezala ulertu zen, hain zuzen *aritmetika politikoa* deitu zitzaion, estatistika izena XIX. mendean nagusitu arte. XIX. mendean, estatistikaren eremua arlo geografiko eta ekonomikotik beste arlo batzuetara zabaldu zuen, gizarte zein natur zientzietara. Hain zuzen, Adolphe Queteletek *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 aztertzeko, baina datuak aztertzeko tresna gisa izan zezakeen balioaz ohartu gabe. XIX. mendean, ordea, estatistika probabilitatearen kontzeptua barneratzen joan, era matematikoan, eta hala metodo estatistikoek zehaztasun matematikoa eskuratzen joan ziren. Bide horretatik XX. mendearen erdialderako estatistika matematikoaren funtsa ia guztiz garatuta geratu zen.

XX. mendearen bigarren erdialdean informatikaren garapena gertatu zen, estatistikari bide berriak zabaldu zizkiona, bereziki datu-multzo handiak aztertzeko. Garai horretan garatzen da *aldagai anitzeko analisia*, aldagai askotako datuak batera aztertzteko metodoak garatzen dituena. 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* deitzen zaio.

1.6 Estatistikaren aplikazio-arloak

Jakintzaren arlo guztietan aplika daiteke estatistika, astronomiatik (izar mota ezberdinen ezaugarriak aztertzeko) literatura (idazle ezberdinen estiloak era kuantitatiboan alderatzeko). Horren froga garbia estatistika unibertsitateko gradu gehienetako ikasketa-planetan 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**, makroekonomiara 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);
- **geoestatistika**, meteorologia, klimatologia, geologia eta beste lurraren zientzietara aplikatzen dena eta bereziki denborazko eta espaziozko datuak aztertzen dituena.

1.7 Ikerketa estatistikoaren plangintza

Datuak modu egokian jasotzen hasi aurretik, garrantzitsua ikerketaren hainbat alderdi zehaztea, zein datu bildu eta nola jaso behar diren erabakitzeko: 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, *behaketa-unitatea*, zeinen gainean jasoko den datu bakoitza alegia, *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* zer den, behaketa-unitatea alegia, zehaztu beharko da; orobat zer esan nahi dugun *Gipuzkoako* esaten dugunean (Gipuzkoako erroldan egotea, Gipuzkoan lan egitea,...), espazio-eremua alegia; eta 2015eko errenta ezartzean zer adierazi nahi den zehazki, aldagaia (errenta) eta denbora-eremua (2015). Puntu 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.

1.8 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 kategoria bakarrik jasotzen dituztenak (adibidez, sexua: gizon/emakume);
  - **aldagai ordinalak**, kategoria ezberdinak mailakatu daitezkeenean; adibidez, ikaslearen kalifikazioa (eskas, nahiko, ...).
- **aldagai kuantitatiboak**, balio kuantitatibo batez *neurtu* egiten dutenak (adibidez, adina urtetan). Aldagai kuantitatiboek zenbat? galderari erantzuten diote.

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