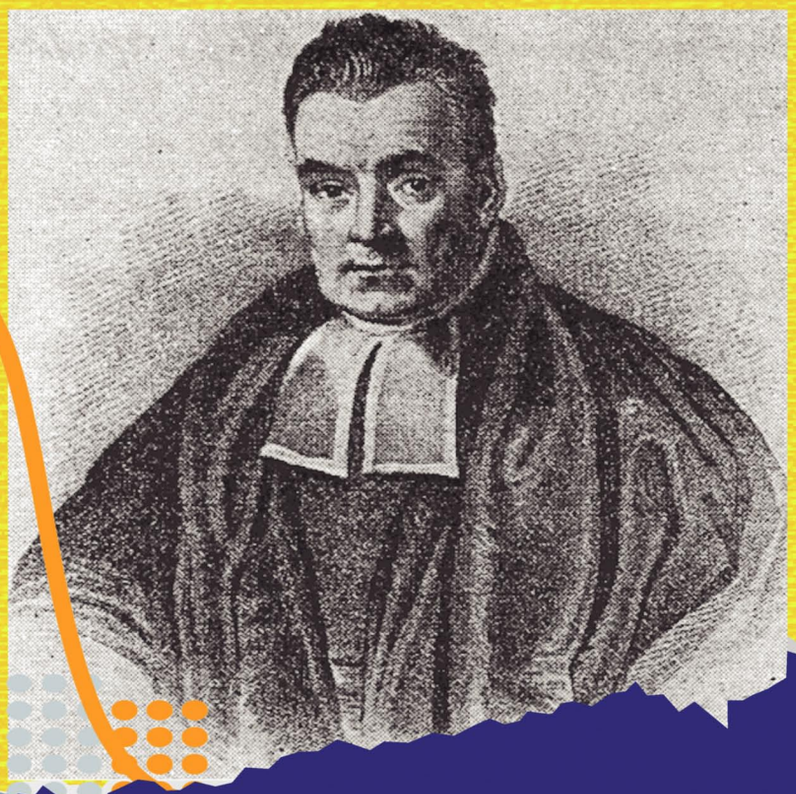


Metodo Bayestarrak Estatistika Ofizialean

2017

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus

Erakunde Autonomiaduna
EUSKO JAURLARITZA
GOBIERNO VASCO

Lanketa:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística

Argitalpena:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística
Donostia-San Sebastián 1
01010 Vitoria-Gasteiz

Ale-kopurua:
Aleak: 100
I/ 2017

Imprimaketa eta Enkoadernazioa:
Eusko Jaurlaritzako Inprimategia eta Erreprografia Zerbitzua

ISBN: 978-84-7749-487-4

Lege-gordailua: VI-603/17

METODO BAYESTARRAK

ESTATISTIKA OFIZIALEAN

Nerea Gutiérrez Aranzabal

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia kalea, 1
01010 VITORIA-GASTEIZ.
Tel.: 945 01 75 00
Faxa: 945 01 75 01
E-mail: eustat@eustat.eus
www.eustat.eus

Aurkezpena

Berriki, Estatistika Bayesiarraren jakinmina Estatistika Ofizialean modu batez nahiko gehitu da. Ildo horretan, Eustat-ek XXVIII Estatistika Nazioarteko Mintegia antolatu zuen izenburu honekin “Calibrated Bayesian Inference for Sample Surveys” (Kalibrazio Bayesiarraren bidezko inferentzia laginketa-inkestetan). Mintegia Roderik Littlek eman zuen, Michigan Unibertsitateko Bioestatistika irakaslea eta Richard D. Remington deritzon katedraduna (EEBB).

Eustat-eko Bekadun batek eremu horretan egindako Ikerketa-lana zabaltzea da argitalpen honen helburua. Dokumentu honek hainbat kapitulu ditu: oinarri teoriko batekin hasiko da, adibide praktikoekin jarraituta, gero aplikazio batekin jarraitzen da non Eustat-eko Informazioaren gizartearen inkesta enpresetan (ESIE) erabiltzen den eredu frekuentista eta eredu bayesiar baten emaitzak alderatzeko eta azkenik, lortutako emaitzak erakusten dira.

Vitoria-Gasteiz, 2017ko martxoak

Josu Iradi Arrieta

EUSTAT-eko Zuzendari Nagusia

Aurkibidea

Atarikoa

Euskal Estatistika Erakundeak (EUSTAT) 2015. urtean estatistika- eta matematika-metodologietan prestatzeko eta ikertzeko emandako bekari esker inferentzia bayestarraren inguruan egin den lanaren emaitza da Koaderno Tekniko honetan bildutakoa.

Dokumentu hau honako kapitulu hauetan banatuta dago:

Lehen kapituluan, inferentzia bayestarraren sarrera moduko bat egiten da, eta joera berri honen eta inferentzia frekuentistaren ezaugarri nagusiak berrikusten dira, haien arteko desberdintasunak azpimarratuta.

Bigarren, hirugarren, laugarren eta bosgarren kapituluetan, hurrenez hurren, proportzio bat, batezbesteko normal bat, batezbestekoen diferentzia eta erregresio lineala kalkulatzeko lantzen da inferentzia bayestarra. Halaber, adibide batzuk ematen dira azaldutako kontzeptu teorikoak hobeto ulertzeko.

Seigarren kapituluan, datu simulatuekin egindako bi adibide ematen dira, eta inferentzia bayestarra Estatistika Ofizialean aplikatzeko aukeraren inguruko lehen ondorioak lortzen dira.

Zazpigarren kapituluan, ikusitako kontzeptuei lotutako aplikazio bat azaltzen da. Kapitulu horretan, EUSTATek egindako Informazioaren Gizarteari buruzko Inkesta – Enpresak (IGIE) erabiltzen da, eredu frekuentista batetik eta eredu bayestar batetik abiatuta merkataritza elektronikoaren estimazioak konparatzeko. Aplikazio horrekin lortutako emaitzak poster-saio baten bidez aurkeztu ziren Toledon, 2016ko irailaren 5etik 7ra bitarte egindako Estatistika eta Ikerketa Operatiboaren eta Estatistika Publikoko X. Jardunaldien XXXVI. Biltzar Nazionalean (SEIO 2016).

Azkenik, 6. eta 7. kapituluen alde praktikoa egiteko R lengoaian inplementatutako kodearen zati bat agertzen duten bi eranskin daude.

Halaber, atariko hau baliatu nahi dut Eustateko Metodologia, Berrikuntza eta I+Gko Arloa osatzen duten guztien babesa eskertzeko, eta, bereziki, Anjeles Iztuetari eskerrak eman nahi dizkiot une oro agertutako babes eta konfiantza irmoagatik. Era berean, Eustateko langile guztien adeitasuna eskertu nahi dut. Izan ere, haiei esker, lan-munduarekiko lehen kontaktu hau esperientzia atsegina eta gogobetegarria izan da, eta betiko iraungo du nire oroimenean. Azkenik, eskerrak eman nahi dizkiet nire familiari eta Gianfeliceri, hartutako erabaki bakoitzean lagundu izanagatik eta une oro nire ondoan egon izanagatik.

GAKO-HITZAK: inferentzia bayestarra, Bayes, inferentzia frekuentista.

1. Estatistika frekuentista vs bayestarra

1.1 Sarrera

"Probabilitate" kontzeptutik abiatuta, bi ikuspegi planteatzen dira analisi estatistikoan: estatistika frekuentista eta estatistika bayestarra. Estatistika bayestarra oraingo kontua dela dirudi, baina urte askotako historia dauka dagoeneko. Izan ere, Bayesen teorema 1763. urtean formulatu zuen Thomas Bayesek. Dena dela, urte asko behar izan ziren paradigma estatistiko horrek eskaintzen dituen abantailen inguruko kontzientzia zientifikoa finkatzeko. Horren arrazoia software bayestarrek izan duten bilakaera izan daiteke. Izan ere, orain konplexutasun handiko problemak ebatz ditzakete, baina garai batean erabat ezinezkoa zitzaien.

1.2 Estatistika frekuentista

Ikuspegi frekuentista batetik, probabilitatea esperimentazioari lotuta definitzen da, eta beti zirkunstantzia beretan errepikatutako proba berdin eta independenteen sekuentzia infinitu bateko maiztasun erlatiboaren muga dela esaten da. Adibidez, esperimentu bat aldi kopuru infinitu batean errepikatzen bada eta 1000 alditik 350etan emaitza jakin bat gertatu dela egiaztatzen bada, frekuentista batek esango luke emaitza horren probabilitatea % 35 dela. Hau da, A gertaera ezezagun baten P probabilitatea, $\mathrm{P(A)}$, gertaera horren maiztasunaren arabera definitzen da, aurreko behaketetan oinarrituta. Definizio horren arabera, frekuentista batek dio gertaera bakoitzari bere egiazko balio bat lortzeko probabilitate bat lotu dakiokeela. Adibidez, Erresuma Batuan jaioberri guztien % 50,9 neskak badira, eta kontuan hartuta A=”ausaz hautatutako ume bat neska da”, $\mathrm{P(A)}$=0,509 izango da.

Beraz, hurbilketa klasikoa gauzen probabilitate "erreala" aztertzean datza, eta esperimentu multzo batekin egindako neurketa bat azpiko probabilitate errealera zer neurritan hurbiltzen den finkatzen saiatzen da. Hori dela-eta, parametroak balio finkoak baina ezezagunak dira frekuentista batentzat, eta datuak behatzeko probabilitatea maximizatzen duten parametroen balioen aukeraketan oinarritzen da estimazioa. Hortaz, ausazko aldagai bati edo gertaera bati lotutako probabilitateaz hitz egin dezakegu, baina inoiz ez parametro, hipotesi, eredu edo falta den behaketa bati lotutako probabilitateaz.

1.2.1 Inferentzia

Populazio baten lagin batetik abiatuta populazio horren inguruko ondorioak ateratzeko aukera ematen diguten estatistika-tekniken multzoa da inferentzia estatistikoa. Ikuspegi frekuentistaren arabera, inferentzia estatistikoko tekniken helburua intereseko parametroari buruzko bi hipotesi kontrastatzea da. Hipotesiak honako hauek dira: $\mathrm{H_0}$ hipotesi nulua eta $\mathrm{H_1}$ hipotesi alternatiboa. $\mathrm{H_0}$ hipotesia era arbitrarioan finkatutako p-balio baten arabera baztertu edo onartuko dugu ($\mathrm{p}<0.05$ normalean). Erabakia hartzeko garaian, bi errore mota egin ditzakegu. Hipotesi nulua bazter dezakegu, hipotesi hori egiazkoa izanik ere (I. motako errorea edo α errorea), edota hipotesi alternatiboa bazter dezakegu, egiazkoa izanik ere (II. motako errorea edo β errorea). II motako errorearen $(1 - \beta)$ probabilitate osagarria kontrastearen ahalmena edo potentzia izenekoa da. Horrela, beraz, I. motako errore bat onartzeko

prestasuna finkatzen dugu. Hala ere, p-ren ezaugarri gisa azpimarra dezakegu bere balioa laginaren tamainaren araberakoa dela. Hau da, lagina zenbat eta handiagoa izan, orduan eta handiagoa izango da konparatzen diren taldeen artean alde esanguratsuak aurkitzeko probabilitatea.

1.2.2 Konfiantza-tarteak

Inferentzia frekuentistan, konfiantza-tarteak (KT) ere oso maiz erabiltzen dira. $IC_{95\%}$ bat aldien $\%\,95$etan asmatzen duen balio-heina da. Hau da, laginketa askotan errepikatuko balitz, prozedura bera aplikatuko balitz eta konfiantza-tarteak $\%\,95$ean kalkulatuko balira (formula ezagunen arabera), 100 tartetik 95etan kalkulatzen ari den egiazko parametroa egongo litzateke, eta α=0,05 dela finkatuko litzateke. Hortaz, konfiantza-tarte bakarra kalkulatzen bada, tarte hori "arrakastatsuetako" bat delako "konfiantza" baino ezin da eduki. Konfiantza-tarteak egiteko garaian, estimazioak eta errore estandarrak esku hartzen dute. Errore estandarra gure estimazioan onartzen dugun errore-maila da, eta laginaren tamainaren alderantziz proportzionala da. Hortaz, lagina zenbat eta handiagoa izan, orduan eta zehatzagoa izango da estimazioa. Horrela, beraz, H_0 kontrastatzea gure konfiantza-tarteak parametroa biltzen duela egiaztatzea da.

1.2.3 Adibidea: proportzioen konparazioa hainbat egoeratan, ikuspegi frekuentista batetik abiatuta

Demagun, adibidez, gaixotasun bat duten pazienteak tratamendu konbinatu bat jasotzen dutenean tratamendu tradizionala erabiltzen dutenean baino bizkorrago osatzen diren baloratzeko saiakuntza kliniko bat egiten dela. Ziur samar dakigu pazienteen $\%\,60$ baino ez direla azken baliabide terapeutikoarekin osatzen. Bestalde, ikerketa-taldeak bibliografian eta eguneroko lanean oinarrituta dituen arrazoi teorikoak eta zantzu enpirikoak kontuan hartuta, baikortasun handi samarrarekin pentsa daiteke tratamendu konbinatua tratamendu sinplea baino eraginkorragoa dela. Demagun n = 80 dela, tratamendu konbinatua 40 pertsonari eta tratamendu tradizionala beste 40ri aplikatzen diegula, eta bost egunen ondoren tratamendu konbinatua jaso duten pazienteen $\%\,75$ek ($\pi_c = 0{,}75$) nabarmen egin dutela hobera, eta tratamendu tradizionala jaso dutenen kasuan, berriz, osatu direnen tasa $\%\,60$ ($\pi_t = 0{,}60$) dela.

	BAI	EZ	Guztira
K	30	10	40
T	24	16	40
Guztira	54	26	80

Datuak lortu ondoren (π_c, π_t y $\pi_d = \pi_c - \pi_t$), honako hipotesi honi lotutako probabilitatea kalkulatzen da: H_0: $\pi_d \leq 0$. Horretarako, χ^2 proba aplikatu, eta $\chi^2{}_{obs} = 2{,}05$ dela eta dagokion p-balioa 0,076 dela lortuko dugu. Balio hori behar bezain txikia ez denez gero, ez dugu ebidentzia nahikoa tratamendu konbinatuak ondorio nabarmen desberdinak dituela ondorioztatzeko.

Demagun esperimentua berriz egiten dela, baina oraingoan n ez dela 40 paziente talde bakoitzean, 200 baizik. Honako emaitza hauek lortuko lirateke:

	BAI	EZ	Guztira
K	150	50	200
T	120	80	200
Guztira	270	130	400

Osatzen diren pazienteen tasak $\pi_c = 0{,}75$ eta $\pi_t = 0{,}6$ dira berriz, eta, beraz, diferentzia $\pi_d = 0{,}15$ da. Kasu horretan, χ^2 probarekin $\chi^2_{obs} = 10{,}26$ lortuko dugu, eta, beraz, p-balioa 0,00068 izango da. Beraz, tratamendu konbinatua eraginkorragoa dela onartuko dugu, $\%\ 95$eko konfiantzarekin (baita $\%\ 99$ko konfiantzarekin ere). Horrek esan nahi du laginaren tamainak eragin handia duela p-balioan.

Azkenik, demagun orain bigarren esperimentuaren baldintza beretan egoera alderantzizkoa dela. Honako emaitza hauek lortuko genituzke:

	BAI	EZ	Guztira
K	103	97	200
T	120	80	200
Guztira	223	117	400

Lortuko genituzkeen estimazioak $\pi_c = 0{,}52$ eta $\pi_t = 0{,}60$ izango lirateke ($\pi_d = \pi_t - \pi_c$ izanik). H_0 hipotesia ($\pi_d \geq 0$) egiazkoa dela suposatuta, hipotesi-probak esaten digu $\chi^2_{obs} = 3{,}29$ eta p-balioa$=0{,}0349$ dela. Kasu horretan, H_0 hipotesia baztertuko dugu, baina kontrako zentzuan, hau da, tratamendu tradizionala konbinatua baino eraginkorragoa dela ondorioztatuko dugu. Ondorio hori ez dator bat aldez aurretiko ezagutzekin eta esperimentuaren bidez bilatzen zen emaitzarekin.

1.3 Estatistika bayestarra

Maiz egiten dira honelako galderak: zenbatekoa da ikasle batek matematikako azterketa gainditzeko probabilitatea? Erantzun bat emateko, inork ez luke beharrezkotzat hartuko ikasleak azterketa 1.000 aldiz egitea, era horretan zenbat aldiz gainditu duen zenbatzeko eta arrakasten ehunekoa kalkulatu ahal izateko. Halaber, ohikoa izaten da "oso handia da Bartzelonak partidua irabazteko probabilitatea" entzutea edota mediku batek paziente jakin bat onik ateratzeko probabilitatea txikia dela esatea. Horrelako baieztapenek ez dute zentzurik esparru frekuentistan, ekintza horiek ezin izango baititugu errepikatu aldi askotan eta baldintza beretan. Lagunarteko hizkeran, berriz, askotan erabiltzen dira, eta zeregin erreal bat betetzen dute erabakiak hartzeko garaian.

Metodo bayestarrak, aldiz, ebidentzia berrietan oinarrituta iragarpenak hobetzeko ideian oinarrituta daude. Ikuspegi bayestarrari jarraikiz, probabilitatea ziurgabetasun-mailaren neurri bat da. Bayestar batentzat, iragarpen bat egiteko dugun ezagutza-maila neurtzen duen adierazpen matematikoa da probabilitatea. Beraz, joera bayestarraren arabera

, zorrotz mintzatuta ez da zuzena “P ekintza bat gertatzeko probabilitatea $\%\ 30$ dela iragartzen dut" esatea. Horren ordez, honako hau adierazi beharko litzateke: "orain daukadan ezagutzan oinarrituta, $\%\ 30$ean ziur naiz P gertatuko dela”.

Bestalde, bayestarrek ausazko aldagai gisa interpretatzen dituzte parametroak, eta ikuspegi honetan funtsezko tresna den Bayesen Teoreman oinarrituta aztertzen da aldagai horien

probabilitate-banaketa. Probabilitate bayestarraren arabera, oinarrizkoagoa da baldintzazko probabilitate bat baterako probabilitate bat baino. Beraz, $P(A|B)$ definitu nahi da $P(A, B)$ baterako probabilitatea kontuan hartu gabe.

1.3.1 Inferentzia bayestarraren lehenengo urratsak

$P(A,B)=P(A|B)*P(B)$ eta, simetriaz, $P(A, B)=P(B|A)*P(A)$ kontuan hartuta, Bayesen formula ezaguna lortzen da bi adierazpenak batuta:

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

Problema bayestar bat formulatu nahi izanez gero, θ deituko diogu, adibidez, kalkulatu nahi dugun parametroari. Ondoren, parametroaren probabilitate-banaketa finkatu behar da eskuragarri dagoen informazioarekin; a priori probabilitate deitzen zaio horri: $P(\theta)$. Oso garrantzitsua da aldez aurretiko informazioa datuak kontuan hartu gabe ezartzea. Estatistika bayestarraren aurkako kritika nagusiak puntu horrengatik jaso dira. Izan ere, objektibotasuna galtzen da, eta, beti esan izan denez, zientziak objektiboa izan behar du. Efektu hori murrizteko, a priori informazioa kuantifikatzeko garaian metodo bakar batera ez mugatzea eta kasu bakoitzean ondorioak nola aldatzen diren egiaztatzea gomendatzen dute egile batzuek. "Sentsibilitate-analisi" deitzen zaio horri. Edonola ere, a priori informazioa nahikoa ez bada eta intuizio pertsonalak alde batera utzi nahi badira, informaziozkoa ez den banaketa bat esleituko da, aurrerago ikusiko dugunez.

Ikuspegi bayestarraren ezaugarri bereizgarri bat zera da, lagin behagarriaren tamaina oso handia denean, aukeratzen den a priori banaketak oso garrantzi txikia izan ohi duela. Horrela, beraz, ikuspegi bayestarraren baliagarritasunik handiena laginen tamainak oso handiak ez direnean lortzen da.

Bigarrenik, gure azterketan behatutako datuek ematen duten informazioa kuantifikatu beharra dago, hots, egiantza: $L_D(\theta)$. Azkenik, a posteriori probabilitate-funtzioa kalkulatzen da: $P(\theta|\text{datos})$. Hau da, zenbateko probabilitatea duten θ parametroaren balioek gure datuak ikusi ondoren. Ondorengo banaketa lortzeko, a priori informazioa eta datuen egiantza konbinatzen dira. Hortaz, $P(\theta|D) \propto L_D(\theta) * P(\theta)$, non $\propto$ sinboloak proportzionaltasuna adierazten duen.

1.3.2 Aplikazioa proba sekuentzialetan

Paradigma bayestarra, beraz, ikaskuntza efektuari lotuta dago. Horren arabera, prozedura dinamiko bati lotuta dago ezagutza. Datu berriak eskuratzean, gure aldez aurretiko ezagutzak datu berriak erantsiz eguneratzeko aukera eskaintzen digu Bayesen teoremak. Horretan oinarrituta, aldez aurretiko informazio berri bihurtzen dira. Hau da, lehenik eta behin, D_1 datu multzoa behatzen da:

$$P(\theta|D_1) \propto L_{D_1}(\theta) * P(\theta).$$

Ondoren, D_2 multzo berria behatu behar da:

$$P(\theta|D_1, D_2) \propto L_{D_2}(\theta) * P(\theta|D_1).$$

Beraz, esan daiteke metodo bayestarrak aproposak direla proba sekuentzialak aztertzeko.

Adibidez, demagun aseguru-konpainia batek, urtero bezala, bere hiru filialetako bakoitzak (Fi_1, Fi_2 eta Fi_3) kobratu beharreko fakturen ikuskapena egin nahi duela. Aurreko ikuskapenetatik dakigunez, fakturen $\%\ 20$ Fi_1 filialarenak, $\%\ 30$ Fi_2 filialarenak eta $\%\ 50$ Fi_3 filialarenak dira. Demagun, halaber, Fi_1 filialaren fakturen $\%\ 1$, Fi_2 filialaren fakturen $\%\ 2$ eta Fi_3 filialaren fakturen $\%\ 3$ akastunak direla. Demagun orain egoitza nagusian faktura guztiak jasotzen direla eta bertan ez daudela batere pozik Fi_3 filialeko zuzendariarekin. Fakturen bolumena oso handia denez gero, multzo bat ausaz hautatzea erabaki da. Demagun ausaz bat hautatzen dugula eta akastuna dela. Faktura hori Fi_3 filialekoa izateko probabilitatea $\%\ 50$etik gorakoa bada, zuzendaria kendu egingo dute kargutik. Helburua hirugarren filialeko zuzendaria kargutik kenduko duten ala ez iragartzea da.

Honako hauek definituko ditugu: F_i=”Hautatutako faktura Fi_i filialekoa da”, i=1,2,3 eta D=”Hautatutako faktura akastuna da”. Asmoa $P(F_3|D)$ kalkulatzea da, eta, horretarako, Bayesen teorema erabiliko dugu honako probabilitate hauekin:

$P(F_3)$: faktura Fi_3 filialekoa izateko a priori edo aldez aurretiko probabilitatea da; F_3 filialaren probabilitatea da, faktura akastuna den ala ez jakin aurretik.

$P(F_3|D)$: a posteriori edo ondorengo probabilitatea; F_3 filialaren probabilitatea da, faktura akastuna dela jakin ondoren. Hau da, Bayesen teoremak hasierako probabilitateak emango dizkigu, behatutako datuekin berrikusiak.

$P(D|F_3)$: D gertaerak (hautatutako faktura akastuna da) duen egiantza, faktura hirugarren filialekoa dela jakinik. Ikusi dugunez, honako notazio hau erabil daiteke:

$$P(F_i|D) \propto L_D(F_i) * P(F_i).$$

Adibide honetan, honako hau dugu: $P(F_1) = 0{,}2$, $P(F_2) = 0{,}3$ y $P(F_3) = 0{,}5$ eta $P(D|F_1) = 0{,}01$, $P(D|F_2) = 0{,}02$ y $P(D|F_3) = 0{,}03$. Hortaz,

$$P(D) = P(D|F_1) * P(F_1) + P(D|F_2) * P(F_2) + P(D|F_3) * P(F_3) = \\ = 0{,}01 * 0{,}2 + 0{,}02 * 0{,}3 + 0{,}03 * 0{,}5 = 0{,}023.$$

Eta

$$P(F_1|D) = \frac{P(D|F_1) * P(F_1)}{P(D)} = \frac{0{,}01 * 0{,}2}{0{,}023} = 0{,}087.$$
$$P(F_2|D) = \frac{P(D|F_2) * P(F_2)}{P(D)} = \frac{0{,}02 * 0{,}3}{0{,}023} = 0{,}261.$$
$$P(F_3|D) = \frac{P(D|F_3) * P(F_3)}{P(D)} = \frac{0{,}03 * 0{,}5}{0{,}023} = 0{,}652.$$

Hau da, Fi_3 filialeko zuzendaria kargutik kenduko dute.

Demagun orain bigarren faktura bat aztertu nahi dugula eta hura ere akastuna dela. Hirugarren filialeko zuzendariaren egoera aztertu behar da berriz.

A priori probabilitate gisa lehen fasean a posteriori gisa lortutakoak erabiliz gero, honako hau izango dugu:

$$P(F_1|D) = 0{,}087;\ P(F_2|D) = 0{,}261;\ P(F_3|D) = 0{,}652;$$

D_2:=”bigarren fasean hautatutako faktura akastuna da" izanik:

$$P(D_2) = P(D_2|F_1) * P(F_1|D) + P(D_2|F_2) * P(F_2|D) + P(D_2|F_3) * P(F_3|D) = \\ = 0{,}01 * 0{,}087 + 0{,}02 * 0{,}261 + 0{,}03 * 0{,}652 = 0{,}02565.$$

Eta

$$P(F_1|D_2) = \frac{P(D_2|F_1) * P(F_1|D)}{P(D_2)} = \frac{0{,}01 * 0{,}087}{0{,}02565} = 0{,}034.$$
$$P(F_2|D_2) = \frac{P(D_2|F_2) * P(F_2|D)}{P(D_2)} = \frac{0{,}02 * 0{,}261}{0{,}02565} = 0{,}2035.$$
$$P(F_3|D_2) = \frac{P(B_2|D_3) * P(F_3|D)}{P(D_2)} = \frac{0{,}03 * 0{,}652}{0{,}02565} = 0{,}7625.$$

Horrela, beraz, egoerak okerrera egin du hirugarren filialeko zuzendariarentzat, faktura bere filialekoa izateko probabilitatea $0{,}7625$ra igo baita.

1.3.3 Sinesgarritasun-tarteak

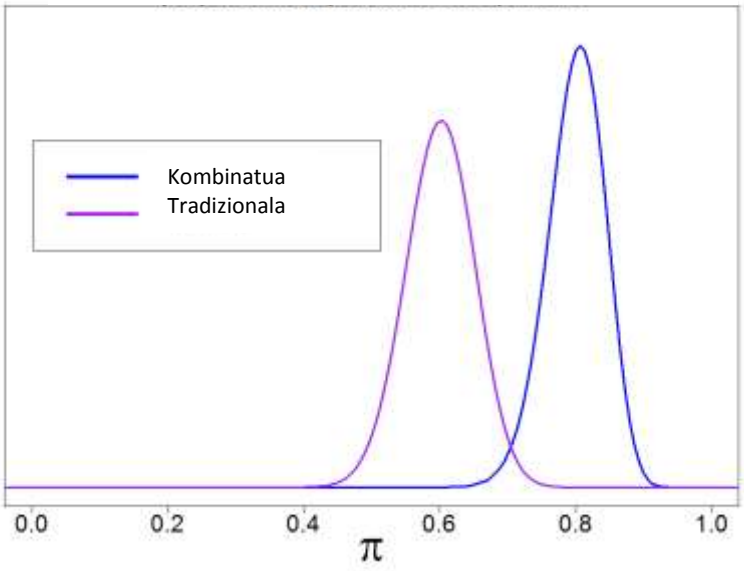
Inferentzia bayestarrean, sinesgarritasun-tarteak erabiltzen dira konfiantza-tarteen ordez. Bayestarrek a posteriori lortzen den dentsitate-funtzioa adierazten duen kurba hartzen dute kontuan, eta kurba horren azpiko azalera X eta Y balioen artean $\%\ (1\text{-}\alpha)$ bada, egiazko balioa probabilitate handiarekin (adibidez $\%\ 95$ekoarekin, baldin eta α=0,05 bada) X eta Y balioen artean dagoela esan daiteke. Kasu horretan, (X, Y) balioek $\%\ 95$eko sinesgarritasun-tartea edo konfiantza-tarte bayestar bat osatzen dutela esaten da. Nolanahi ere, ikuspegi frekuentista batetik eta ikuspegi bayestar batetik emaitza berdinak lor badaitezke ere, emaitza horien interpretazioa desberdina izango da beti. Adibidez, a priori banaketa finkatzeko garaian banaketa hori erabiltzeko behar adinako ebidentziarik ez dugula erabaki badugu, eta, beraz, a priori banaketa gisa banaketa uniformea finkatzen badugu, lortuko ditugun konfiantza-tarteak berdinak izango dira zifrari dagokionez. Ikusi dugunez, berriz, esanahia oso desberdina izango da.

1.3.4 Adibidea: proportzioen konparazioa hainbat egoeratan, ikuspegi bayestar batetik abiatuta

Bestalde, tratamendu konbinatuaren eta tratamendu tradizionalaren arteko aldea esanguratsua zen ala ez aztertu dugun 1.2.3 atalera itzuliz, π_c eta π_t proportzioei buruzko ezagutza handiagoa lortu nahi dugu metodo bayestarrak erabiliz. Bi proportzioak $[0{,}1]$ tartean dauden balio jarraituak dira, eta balio horien balio zehatza zenbatekoa den ez badakigu ere, badugu intuizio bat, eta etekinik handiena atera behar diogu. Ikusi dugunez, gure hasierako ideiak proportzio bakoitzerako probabilitateen banaketa baten bidez adieraztean datza lehen urratsa. Aurreko azterketen ondorioz, zantzu handiak genituen tratamendu konbinatua tratamendu tradizionala baino eraginkorragoa zela pentsatzeko. Hori dela-eta, a priori probabilitate gisa pentsa daiteke π_t balioa seguruenik 0,4tik 0,8rako tartean dagoela, 0,6 inguruko balioa duela eta puntu horretatik urruntzean probabilitatea berehala murrizten dela. π_c proportzioari dagokionez, 0,8 baliotik hurbil dagoela eta (0,7-0,9) tartetik kanpo egoteko probabilitatea urria dela pentsatzen dugu. Proportzio bat denez gero, 2. kapituluan ikusiko dugunez a priori dentsitate baten ohiko forma a eta b parametroen mendeko beta banaketa batena da. Beta dentsitatearen funtzioa honako honen proportzionala da: $\pi^{a-1}(1-\pi)^{b-1}$.

π_c proportzioaren kasuan, a priori beta banaketa $a = 72$ eta $b = 18$ parametroak dituena izan daiteke, eta π_t proportzioaren kasuan, berriz, $a = 57$ eta $b = 38$ parametroak dituena. Ondorengo irudian, hautatutako a priori dentsitatearen funtzioak daude adierazita.

Osatutako pazienteen proportzioaren a priori banaketak



Bistakoa da a priori banaketak beste modu batean ere aukeratu zitezkeela, baina, guztiz bestelako a priori informazioa edukitzearen kasuan izan ezik, emaitzak ez dira oso desberdinak izango.

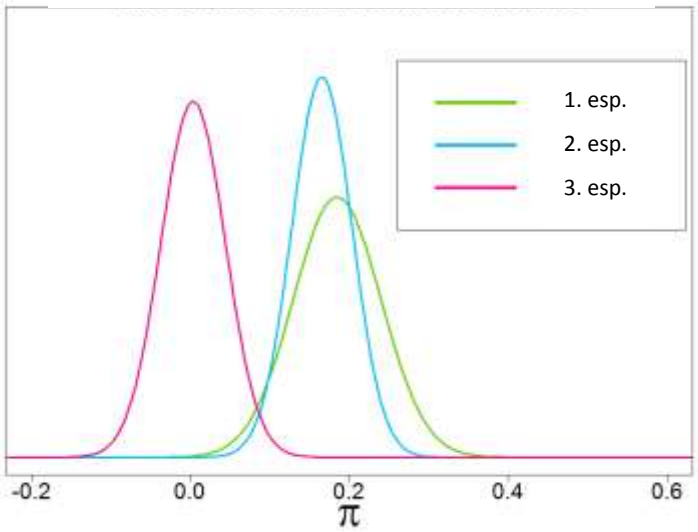
Hurrengo urratsa datuak behatzean datza, eta a posteriori banaketak lortzea izango da helburua. Horretarako, Bayesen teorema erabiliko da. Egiazta daiteke $\text{beta}(a,b)$ banaketa baten eguneratzea (e arrakastak eta f porrotak adierazten dituen binomialaren bidez) $\text{beta}(a^*,b^*)$ banaketa bat dela, $a^* = a + e$ eta $b^* = b + f$ izanik. Lehen esperimenturako, eta lehen taulan agertutako datuak kontuan hartuta, π_c eta π_t proportzioetarako a posteriori banaketak $B_c\,(102{,}28)$ eta $B_t\,(81{,}54)$ izango lirateke, hurrenez hurren.

Gure helburua tratamenduen proportzioek duten diferentziaren ($\pi_d = \pi_c - \pi_t$) banaketa aztertzea da; 4. kapituluan sakonean aztertuko dugunez, banaketa normal bat izango da –$N(m'_d, (s'_d)^2)$–, non

$$m'_d = \frac{a_c^*}{a_c^* + b_c^*} - \frac{a_t^*}{a_t^* + b_t^*}, \quad (s'_d)^2 = \frac{a_c^* b_c^*}{(a_c^* + b_c^*)^2(a_c^* + b_c^* + 1)} + \frac{a_t^* b_t^*}{(a_t^* + b_t^*)^2(a_t^* + b_t^* + 1)}.$$

R softwarea erabiliz, esperimentu bakoitzaren diferentziaren banaketak honako hauek dira:

Osatutako pazienteen proportzioetan dagoen diferentziaren a posteriori banaketak



Lehen esperimentuaren banaketari erreparatzen badiogu, diferentziaren eremurik probableena 0,18 balioaren ingurukoa dela ikusiko dugu. Gogora dezagun ikuspegi frekuentistaren barruan ikertzaileak ezin zuela ezer ondorioztatu, χ^2 testa egin ondoren tratamendu konbinatua hobea zela esateko lagin-ebidentzia nahikoa ez zegoela baino ezin baitzuen baieztatu. Ikuspegi bayestarrak, aldiz, baieztapen zehatzak eta zentzuzkoak egiteko aukera ematen duten probabilitateak zenbatzea ahalbidetzen du. Kasu honetan, adibidez, adierazitako tratamenduekin sendatzeko proportzioen arteko diferentzia gutxienez $\%\,10$ izateko probabilitatea $\%\,93{,}7$ da. Hau da, esan daiteke ia ziurra dela tratamendu konbinatua eraginkorragoa dela.

Bigarren esperimentuko datuekin lortutako kurbak agerian uzten duenez, laginaren tamainak analisi bayestarrean duen eragina ez da ikuspegi frekuentistan duena bezain handia. Kasu honetan, laginaren tamaina aurreko kasuan baino askoz handiagoa bada ere, ateratzen dugun ondorioa antzeko samarra da. Izan ere, sendatzeko proportzioen arteko diferentzia 0,17 da kasu honetan, eta 0,18 aurrekoan. Oraingoan, kurba estuagoa da, eta, beraz, laginaren tamaina horrekin diferentziaren portaera posiblearen inguruan lortzen den ideia zehatzagoa da. Hortaz, esan dezakegu tratamendu konbinatuaren balioa balio tradizionala baino gutxienez $\%\,10$ handiagoa izateko probabilitatea $\%\,95{,}85$ dela.

Azkenik, hirugarren esperimentuan, lortutako emaitzak ez datoz bat ikertzaileak aldez aurretik uste zuenarekin. Irudian ikus daitekeenez, tratamenduen arteko diferentzia zero ingurukoa da. Kontraste frekuentistaren arabera, eta $\%\,95$eko konfiantzarekin, tratamendu tradizionala eraginkorragoa zela onartzen zen. Ikuspegi bayestarraren arabera, berriz, tratamendu tradizionalarekin osatzen direnen ehunekoa $\%\,10$etik gorakoa izateko probabilitatea ia mespretxagarria da, eta osatzen direnen ehunekoak 0ren desberdinak gehienez ere $\%\,5$ean izateko probabilitatea $\%\,87{,}53$ra iristen da. Kasu horretan, ikertzaileak egindako a priori probabilitateak bereziki erakargarriak dira: emaitza enpirikoa bere aurreko ezagutzen eta esperientzien guztiz kontrakoa bada ere, lortzen den ondorioa bitartekoa da eta ez du laginean lortutako emaitzak soilik ordezten.

Hortaz, ondoriozta dezakegunez, metodo frekuentistek askotan ematen dizkigutenak baino ondorio intuitiboagoak eta zentzu komunari lotuagoak lor ditzakegu metodo bayestarrekin. Eta

metodo klasikoak aplikatzea askoz errazagoa bada ere, azterketa egin aurreko ikuspuntuak edo usteak txertatzeko aukera izatea oso erakargarria da, oso prozesu naturala eta giza arrazoibidearen antzekoa baita.

1.4 Paradigma bakoitzaren ezaugarri nagusien laburpena

Kapitulu hau amaitzeko eta lehen kontzeptuak finkatzeko, ikuspegi bakoitza definitzen duten ezaugarri garrantzitsuak azalduko ditugu orain.

1.4.1 Estatistika frekuentistaren ezaugarriak

- Ibilbide luzea du egina, eta prozedura eta emaitza onak eskaintzen ditu.
- Erraz aplikatzen da konputazionalki.
- Probabilitatea zirkunstantzia beretan errepikatutako proba berdin eta independenteen sekuentzia infinitu bateko maiztasun erlatiboaren muga dela esaten da.
- Gertaera bakoitzari bere egiazko balio bat lotzeko probabilitate bat eman dakioke.
- Parametroak finkoak baina ezezagunak dira.
- Estimazioa, nolanahi ere, datuak behatzeko probabilitatea maximizatzen duten parametroen balioen aukeraketan oinarrituta dago.
- Egiantz handienaren edo karratu txikienen metodoa ohikoa izaten da.
- Inferentzia hipotesi bat α bat finkatuz kontrastatzean datza.
- H_0 onartzen da, baldin eta $P(\text{datuak}|H_0) > \alpha$ bada.
- H_1 onartzen da, baldin eta $P(\text{datuak}|H_0) < \alpha$ bada.
- P-balioa gutxienez benetan lortu dena bezain muturrekoa izango den emaitza bat lortzeko probabilitatea da, hipotesi nulua egiazkoa dela suposatuta.
- Laginaren tamainak ez du eraginik p-balioan.
- Tamaina bereko hainbat lagin ateratzen badira eta lagin bakoitzaren $\%\,95$eko konfiantza-tarteak kalkulatzen badira, tarte guztien $\%\,95$ek zenbatesten ari den parametroa edukiko dute.
- Aldez aurretiko ezagutzak ez dira kontuan hartzen.

1.4.2 Estatistika bayestarraren ezaugarriak

- Oso ikaskuntza-prozesu naturala da, giza arrazoibidearen antzekoa, eta ondorio intuitiboagoak ateratzeko aukera ematen du.
- Konputazio-denbora gehiago eta algoritmo konplexuagoak behar dira.

- Oso erabilgarria laginen tamaina oso handia ez denean.
- Behatutako ebidentzia berrietatik abiatuta iragarpenak hobetzea da helburua.
- Probabilitatea ziurgabetasun-mailaren neurritzat hartzen da.
- Parametroak ausazko aldagai gisa interpretatzen dira, eta Bayesen teoreman oinarrituta aztertzen da aldagai horien probabilitate-banaketa.
- Aldez aurretiko informazioa erabilgarria da, baina garrantzitsua da aldez aurretiko banaketa behatutako datuak kontuan hartu gabe ezartzea.
- A priori daukagun informazioa kuantifikatzeko eta ondorioak konparatzeko hainbat metodo erabiltzea gomendatzen da, objektibotasuna gal ez dadin.
- Laginaren tamaina oso handia denean, a priori banaketak oso garrantzi txikia izan ohi du.
- Egokia da proba sekuentzialak egiteko.
- % 95eko probabilitatea dago sinesgarritasun-tarteak edo konfiantza-tarteak parametroa edukitzeko.

2. Proportzio baterako inferentzia bayestarra

2.1 Sarrera

Inferentzian gehien agertzen diren arazoetako bat π populazio-proportzio baten balio ezezaguna zenbatestea da. Demagun populazioaren ausazko lagin bat dugula, Yn saiakuntza independenteren sekuentzia bateko arrakasta kopurua dela eta erantzun-aldagaia dikotomikoa dela, hau da, gehienez ere bi balio har ditzakeela. Horietako bati arrakasta (e, "éxito" hitzetik) deitzen zaio, eta gertatzeko probabilitatea π da; besteari, berriz, porrot (f, "fracaso" hitzetik) deitzen zaio, eta gertatzeko probabilitatea 1 – π da.

Y behaketaren (hots, gertaeren guztizkoaren, π parametroa emanda) baldintzazko banaketa $Binomial(n, \pi)$ da, eta honela adierazten da:

$$f(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 1, \dots, n.$$

Kontuan hartu behar da π parametro finko bat dela hemen, eta y-ren balio posibleen banaketaren probabilitatea behatzen ari garela.

Dena dela, y-ren eta π-ren arteko erlazioa aztertzean y parametro finkoa bada eta π har ditzakeen balioen artean uzten badugu, lortzen den egiantz-funtzioa honako ekuazio honen bidez adieraz daiteke:

$$f(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad 0 \leq \pi \leq 1.$$

Lehen begiratuan adierazpen matematiko berdina ikusten bada ere, interpretazioa guztiz desberdina da.

Bayesen Teorema aplikatzeko, datuak behatu aurretik genituen ezagutza eta usteekin eraikitako a priori banaketa-funtzioa –$g(\pi)$– behar dugula ikusi dugu arestian. A posteriori banaketa, beraz, a priori banaketa eta egiantza biderkatuta lortzen da. Biderkadura horrek zentzua dauka baldin eta soilik baldin aldez aurretiko banaketa egiantzarekiko independentea bada. Horregatik, funtsezkoa da aldez aurretiko banaketak behatutako datuen eragina ez izatea. Kasu horretan, a posteriori banaketa biderkadura horren proportzionala izango da.

$$g(\pi|y) \propto g(\pi) * f(y|\pi). \qquad (2.1)$$

Proportzionala dela esaten da, biderkadura hori kurbaren azpiko azalera 1 dela ziurtatuko digun k konstante batekin zatitu behar dugulako dentsitate-funtzio zehatza lortzeko. k hori honako integral honen bidez lortzen da: $g(\pi) * f(y|\pi)$. Hortaz,

$$g(\pi|y) = \frac{g(\pi) * f(y \mid \pi)}{\int_0^1 g(\pi) * f(y \mid \pi)}. \qquad (2.2)$$

Lehen begiratuan, izendatzailearen integralak arazoak sortuko dizkigula dirudi, zenbakizko kalkulu asko egin behar baitira balioa lortzeko. Ikusiko dugunez, berriz, dentsitate-funtzioen propietateak baliatuta, askotan ez dugu urrats hori egin behar izango.

2.2 Aldez aurretiko banaketa uniformea

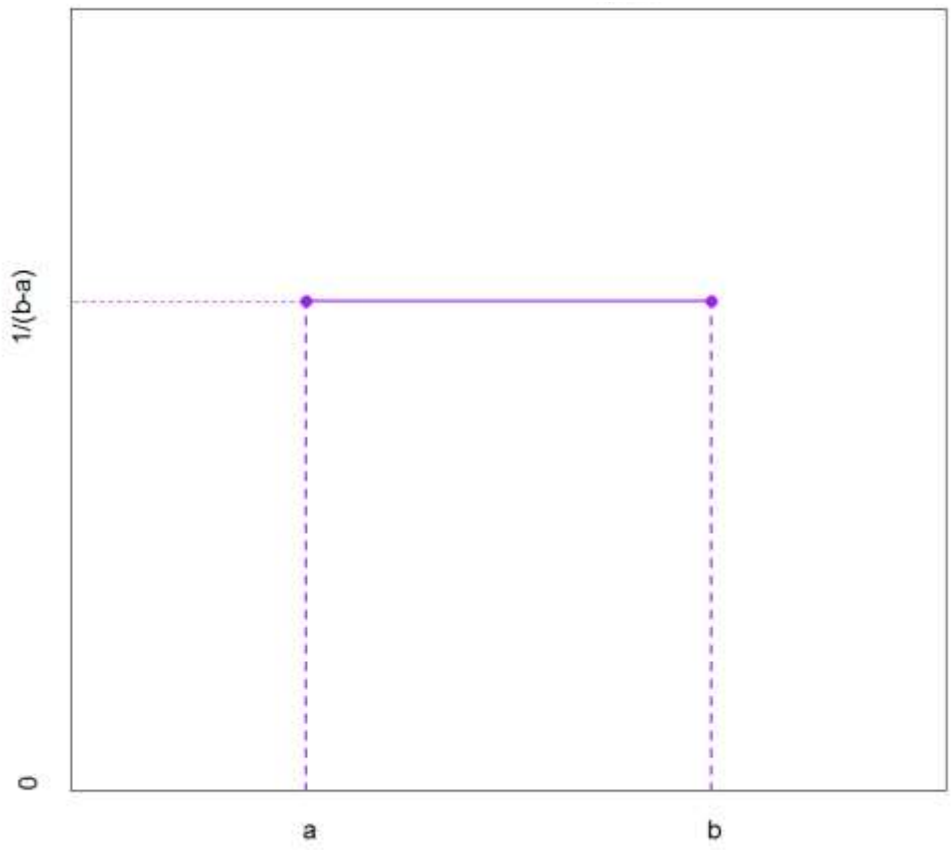
Banaketa uniforme jarraitua ausazko aldagai jarraituetarako probabilitate-banaketen "familia" bat da, eta, banaketa horretan, ausazko aldagaiak probabilitate berdinarekin hartzen du balio bakoitza. Eremua a eta b parametroen bidez definituta dago, eta parametro horiek balio minimo eta maximoak dira. Banaketa forma laburtuan adierazten da maiz: $\mathrm{U(a,b)}$.

Banaketa uniforme jarraituaren probabilitate-dentsitatearen funtzioa honako hau da:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & x < a \text{ o } x > b \end{cases}$$

Eta haren irudikapena, berriz, honako hau:

U(a,b) banaketa baten dentsitate-funtzioa



Erraz ondoriozta daiteke banaketa uniforme baten batezbestekoa, mediana eta bariantza honela adieraz daitezkeela:

$$m = \frac{a+b}{2}, \qquad Mediana = \frac{a+b}{2}, \qquad Var = \frac{(b-a)^2}{12}.$$

Baldin eta $a=0$ eta $b=1$ bada, emaitzazko banaketari ($\mathrm{U}(0,1)$) banaketa uniforme estandar deitzen zaio, eta $\mathrm{beta}(1,1)$. banaketaren baliokidea da. Bestalde, aldez aurretiko behar adina ezagutzarik ez dugunean, $\mathrm{U}(0,1)$ banaketa da nagusi inferentzia bayestarrean. Hortaz, datuek eskaintzen diguten informazioan soilik oinarritutako emaitzak lortu nahi baditugu, π-ren balio posible guztiei pisu bera emango dien dentsitate-funtzio uniforme estandar bat erabili beharko dugu; hau da:

$$g(\pi) = 1, \qquad 0 \leq \pi \leq 1.$$

Kasu horretan, nabaria da a posteriori banaketa egiantzaren proportzionala dela:

$$g(\pi|\mathrm{y}) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad 0 \leq \pi \leq 1.$$

π-ren araberakoa ez den partea baztertu egin daiteke. Izan ere, π-ren balioa edozein izanik ere beti egongo den konstante bat da, eta, beraz, ez du eraginik a posteriori banaketan. $\binom{\mathrm{n}}{\mathrm{y}}$ ezabatzean, $\mathrm{beta}(\mathrm{a}^*, \mathrm{b}^*)$ banaketa ikus dezakegu, non $\mathrm{a}^* = \mathrm{y} + 1$ eta $\mathrm{b}^* = \mathrm{n} - \mathrm{y} + 1$, edota, era baliokidean, $\mathrm{a}^* = \mathrm{a} + \mathrm{e}$ eta $\mathrm{b}^* = \mathrm{b} + \mathrm{f}$. Beraz, erraza da π-ren a posteriori banaketa lortzea integralik egin behar izan gabe.

2.3 Aldez aurretiko beta banaketa

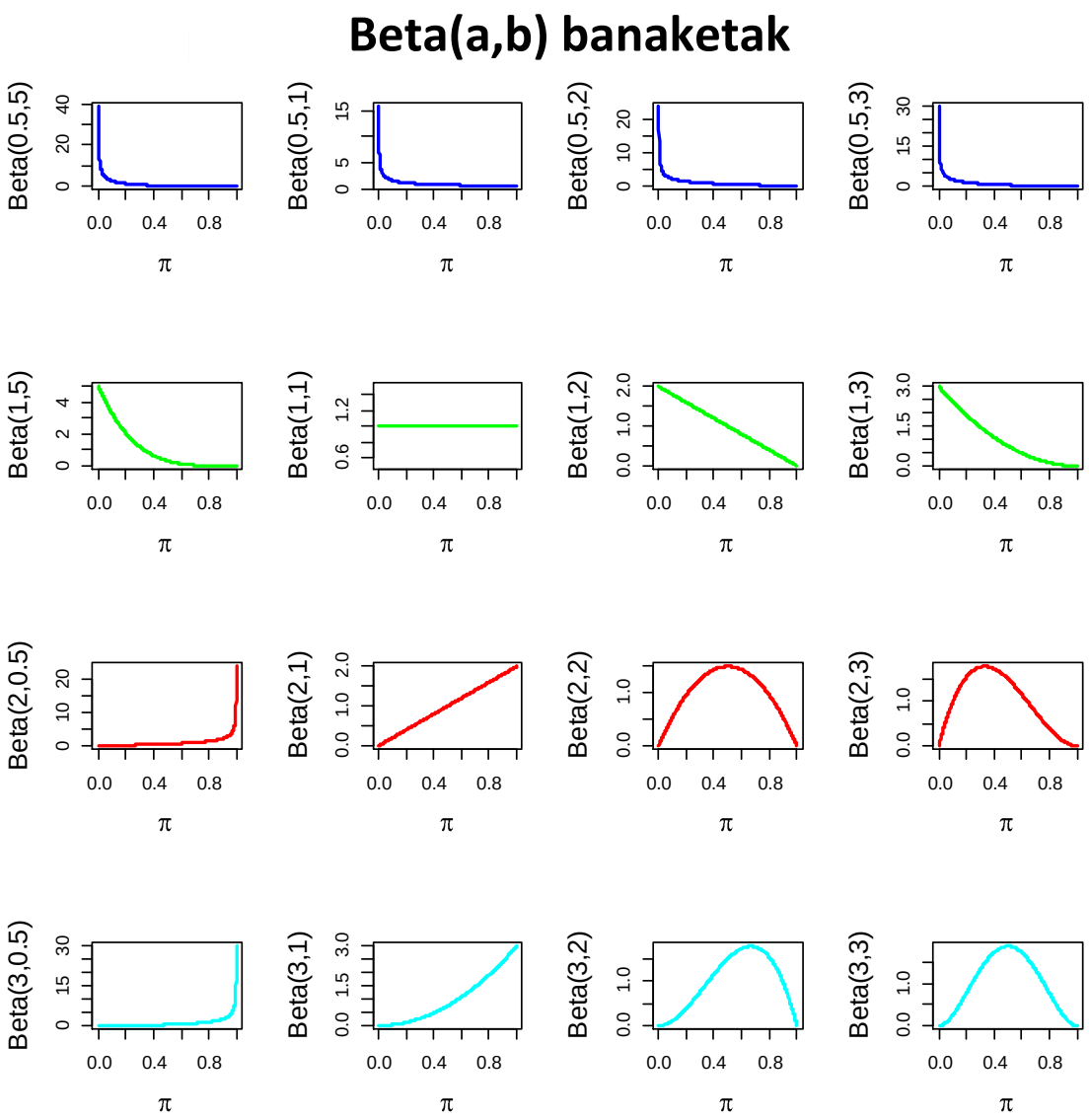
Askotan, balioak $[0,1]$ tartean dituzten ausazko aldagai jarraituekin erabiltzen da beta banaketa, eta, horregatik, oso egokia da proportzioak modelatzeko. Inferentzia bayestarrean, a priori banaketa gisa erabiltzen da maiz, behaketek banaketa binomiala dutenean.

Dentsitate-funtzioa honela adierazten da, $\Gamma(c)$ Gamma funtzioa izanik:

$$f(x;a,b) = \begin{cases} \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} & 0 \leq x \leq 1 \\ 0 & x < 0 \text{ o } x > 1 \end{cases}$$

Banaketa horren alderdi positibo nagusietako bat banaketa enpiriko ugaritara doitzeko aukera da, oso forma desberdinak har baititzake banaketa definitzen duten a eta b parametroen balioen arabera.

Ondorengo beta banaketen adibideetan, a eta b parametroek honako balio hauek hartzen dituzte: $\{0.5, 1, 2, 3\}$. Ikus daitekeenez, $\mathrm{a} < \mathrm{b}$ denean dentsitateak pisu handiagoa du ezkerraldean, eta $\mathrm{a} > \mathrm{b}$ denean, berriz, eskuinaldean; $\mathrm{a} = \mathrm{b}$ denean, funtzioa simetrikoa da. Azkenik, seigarren grafikoan argi eta garbi ikus daiteke $beta(1,1)$ banaketaren funtzioa banaketa uniformearen baliokidea dela [0,1] tartean.



Dakigunez, beta banaketa baten batezbestekoa, moda eta bariantza honela adierazten dira:

$$m = \frac{a}{a+b}, \qquad Moda = \frac{a-1}{a+b-2}, \qquad Var = \frac{ab}{(a+b)^2(a+b+1)}.$$

Inferentzia bayestarrera itzuliz, π-rako erabilitako a priori banaketa-funtzioa $\mathrm{beta}(\mathrm{a}, \mathrm{b})$ izanik:

$$g(\pi; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}, \qquad 0 \le \pi \le 1.$$

Bayesen teoremaren emaitzek konstanteekin biderkatzean aldatzen ez direla dakigunez gero, π-ren araberakoak ez diren parametroak baztertu egin daitezke, eta (3.1) ekuazioa izango dugu:

$$g(\pi|y) \propto \pi^{a-1}(1-\pi)^{b-1} * \pi^{y}(1-\pi)^{n-y}, \qquad 0 \le \pi \le 1.$$
$$g(\pi|y) \propto \pi^{a+y-1}(1-\pi)^{b+n-1}, \qquad 0 \le \pi \le 1.$$

Adierazpen horrek ondorengo banaketaren forma deskribatzen du π.-ren funtzio baten bidez. Berriz ere, beta banaketa bat jasotzen dugu $a' = a + y$ eta $b' = b + n - y$ parametroekin. Hau da, arrakasta kopurua a parametroari gehitzen zaio, eta porrot kopurua, berriz, b parametroari. Era horretan, a posteriori banaketa lortzen dugu berriz, integralik egin behar izan gabe:

$$g(\pi|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\pi^{y+a-1}(1-\pi)^{n-y+b-1}, \qquad 0 \leq \pi \leq 1.$$

Horregatik, behaketa binomialak ditugunean komeni da $beta(a,b)$ banaketa bat erabiltzea. Izan ere, kalkuluak sinpleak dira eta, gainera, Bayesen teorema aplikatzean, familia bereko banaketa berri bat lortuko dugu.

2.4 Aldez aurretiko banaketaren aukeraketa

Bayesen teoremak gure aldez aurretiko usteak datuek ematen duten informazioarekin eguneratzeko aukera ematen digu. Horregatik, datuak aztertu aurretik, garrantzitsua da gure ezagutzak era egokian adierazten jakitea.

2.4.1 Aldez aurretiko informazio urriarekin

Aldez aurretiko informazio gutxi dugunean, aurreko irudian proposatutako banaketa bat finka dezakegu a priori banaketa gisa. Adibidez, π-ren balioa oso txikia dela uste badugu, komenigarria izango litzateke honako banaketa hauetako bat erabiltzea: $beta(0.5,1)$, $beta(0.5,2)$, $beta(0.5,3)$, $beta(1,2)$ edo $beta(1,3)$. Guztiek kalkulu errazak eskainiko dizkigute, eta emaitzak antzeko samarrak izango dira kasu guztietan.

2.4.2 Aldez aurretiko informazio errealarekin

Aukeratutako banaketa ahal den neurrian gure ustearekin bat etor dadin, ezagutzen dugunaren batezbesteko eta desbiderapen estandarra erabiltzea iradokitzen da. π_0 proportzioaren a priori batezbestekoa eta σ_0 a priori desbiderapen estandarra izanik.

Ikusi dugunez, $beta(a,b)$ banaketa baten batezbestekoa honela adierazten da: $\pi_0 = \frac{a}{a+b}$; eta desbiderapen estandarra, berriz, honela: $\sigma_0 = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$. Hortaz, $\frac{b}{a+b} = 1 - \pi_0$ izanik, eta desbiderapen estandarrean adierazpen hori eta batezbestekoaren adierazpena ordeztuz gero, honako hau lortuko dugu: $\sigma_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{(a+b+1)}}$. Horrela, beraz, ondorengo sistema ebatzita, gure $beta(a,b)$ banaketaren a eta b parametroak lortuko ditugu.

$$\begin{cases} \pi_0 = \dfrac{a}{a+b} \\ \sigma_0 = \sqrt{\dfrac{\pi_0(1-\pi_0)}{(a+b+1)}} \end{cases} \qquad (2.3)$$

2.4.3 Kontuan hartu beharreko alderdiak

Komeni da aukeratutako banaketa marraztea, balioak arrazoizkoak direla eta gure ezagutzarekin bat datozela egiaztatzeko. Bat ez badatoz, π_0 eta σ_0-ren balioak alda daitezke kurbak nahi dugun forma eduki arte.

Bestalde, komeni da, halaber, aldez aurretikoaren lagin-tamaina baliokidea kalkulatzea. $Binomial(n, \pi)$ baten $\hat{\pi} = \frac{y}{n}$ proportzioak honako bariantza hau du: $\frac{\pi(1-\pi)}{n}$. Bariantza hori aldez aurretiko bariantzarekin berdinduz (a priori batezbestekoaren bidez):

$$\frac{\pi_0\,(1-\pi_0)}{n_{eq}} = \frac{ab}{(a+b)^2(a+b+1)}\,.$$

$\frac{\mathrm{a}}{\mathrm{a+b}} = \pi_0$ eta $\frac{\mathrm{b}}{\mathrm{a+b}} = 1 - \pi_0$ denez gero, lagin-tamaina baliokidea $\mathrm{n_{eq}} = \mathrm{a} + \mathrm{b} + 1$ da. Hau da, aukeratu dugun banaketak ematen duen informazio kantitatea tamaina horretako ausazko lagin baten baliokidea da. Horrela, beraz, aukeratutako banaketa errealista den edo aldez aurretiko banaketa solidoegi batekin lan egiten ari garen egiaztatu ahal izango dugu. Kasu horretan, desbiderapen estandarraren balioa handitu beharko dugu gure a priori ezagutzaren pisua hain handia izan ez dadin.

2.4.4 Aldez aurretiko banaketa jarraitu bat sortzea

Demagun orain gure ustea ez dela behar bezala adierazten $beta(a, b)$ banaketa baten bidez. Kasu horretan, gure ezagutzarekin bat datorren aldez aurretiko diskretu bat sor daiteke, eta, interpolazio bidez, jarraitu bihurtu. Bayesen teoremaren propietateei esker, ez da beharrezkoa funtzioa benetako dentsitate-banaketa bat izatea eragiten duen konstantea sartzea. Dena dela, egiantzaren eta aldez aurretikoaren arteko biderkaduraren integrala kalkulatu beharko da.

2.4.5 Aldez aurretiko banaketaren eragina

Lan egiteko erabiltzen ditugun datuen bolumena zenbat eta handiagoa izan, orduan eta txikiagoa izango da aldez aurretiko banaketaren eragina. Horrela, beraz, aldez aurretiko desberdin samar batzuetatik abiatu bagara ere, oso antzekoak diren a priori banaketak lortu ahal izango ditugu.

2.4.6 Adibidea

Ikastetxe bateko zuzendaria, matematikako irakaslea, ingeleseko irakaslea eta haur baten ama eztabaidatzen ari dira matematikako hurrengo azterketa gaindituko dutenen ehunekoari buruz. Zuzendaria oso baikorra da, bere kargu irakaslerik onenak dituelakoan baitago. Harentzat, a priori proportzioaren batezbestekoa 0,8 da, 0,08ko desbiderapenarekin. Hortaz, zuzendariaren ustearekin bat datorren $beta(a, b)$ banaketaren parametroak lortuko ditugu, eta (2.3) sistema ebatziko dugu:

$$\begin{cases} 0.8 = \dfrac{a}{a+b} \\ 0.08^2 = \dfrac{0.8 * 0.2}{a+b+1} \end{cases}$$

Horrela, beraz, zuzendariarentzat, a=19.2 eta b=4.8 izango da, $a + b + 1 = 25$ lagin-tamaina baliokidea izanik.

Matematikako irakasleak ez du bere ustera ongi hurbiltzen den banaketarik aurkitzen, bere ustea forma trapezoidal batekin hobeto irudikatzen dela pentsatzen baitu. Ondorengo taulan, irakasle horrek balio posible bakoitzari ematen dion pisua jaso dugu.

Balioa	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Pisua	0	0	0	1	2	3	3	3	2	1	0

Balioak linealki interpolatuz, honela adieraz dezakegu irakaslearen aldez aurretiko banaketa:

$$g(\pi) = \begin{cases} 10\pi - 2 & 0.2 \leq \pi < 0.5 \\ 3 & 0.5 \leq \pi < 0.7 \\ -10\pi + 10 & 0.7 \leq \pi \leq 1 \end{cases}$$

Ingeleseko irakasleak ez daki ikasleek nolako trebetasunak dituzten matematikan. Horregatik, ez du kontuan hartzen a priori informaziorik, eta banaketa uniforme bat erabiltzen du. Hortaz, irakasle horrentzat, $a=b=1$ da, eta lagin-tamaina baliokidea, berriz, $a+b+1=3$.

Azkenik, ikaslearen amak eskolara joaten zen garaiko oroitzapenetan oinarritzen du bere ezagutza. Oroitzapen horien arabera, 30 ikasle zeuden amaren ikasgelan, eta gainditu dutenen ehunekoa $\%\,50$ inguru zen. Kasu horretan, honako sistema hau ebatzi behar dugu a priori $beta(a, b)$ banaketaren parametroak lortzeko:

$$\begin{cases} 0.5 = \dfrac{a}{a+b} \\ 30 = a + b + 1 \end{cases}$$

Hortaz, amaren usteak $beta(24.5{,}24.5)$ banaketari jarraitzen dio, eta banaketa hori simetrikoa da a priori batezbestekoarekiko.

Handik egun batzuetara, 100 ikaslek egin dute azterketa, eta 74k gainditu dute ($y=74$). Zuzendariaren, ingeleseko irakaslearen eta amaren ondorengo banaketak automatikoki lortzen dira:

$$beta(19.2 + 74, 4.8 + 26) = beta(93.2, 30.8),$$

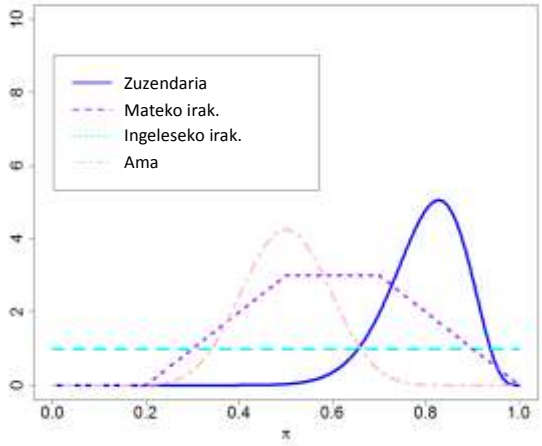
$$beta(1 + 74, 1 + 26) = beta(75, 27),$$

$$beta(24.5 + 74, 24.5 + 26) = beta(98.5, 50.5).$$

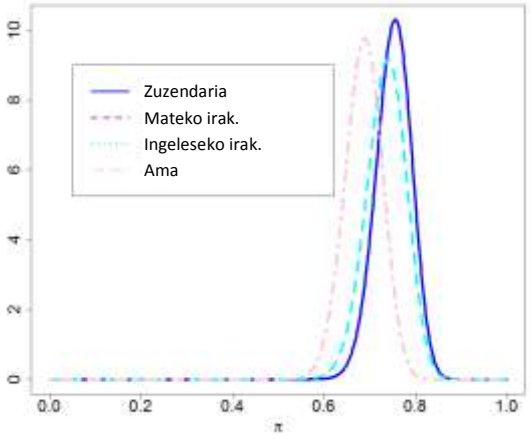
Matematikako irakaslearena lortzeko, berriz, kalkuluren bat egin beharko dugu. Horretarako, R lengoaian inplementatutako Bolstad liburutegia erabiliko dugu.

Ondorengo irudietan, a priori eta a posteriori banaketak ditugu.

Gainditu dutenen proportzioaren aldez aurretiko banaketak



Gainditu dutenen proportzioaren ondorengo banaketak



Lehen hiru kasuetan, a priori banaketa guztiz desberdinekin hasita ere, a posteriori banaketak antzekoak dira. Adibidez, ingeleseko irakasleak gaiari lotutako aldez aurretiko usterik ez zuenez gero, behatutakoaren eragina jasotzen du. Zuzendariak, berriz, 25 ikasleko lagin baten inguruko ezagutzak zituen, eta 100 ikaslerekin lortutako datuak uste baino okerragoak izan direnez gero, zuzendariaren a posteriori banaketak ezkerrera egin du zertxobait. Eta, azkenik, amak 30 ikasleko lagin batean behatutakoarekin justifikatzen zuen bere ezkortasuna. 100 ikasleren azterketen emaitzak ikusi ondoren, gainditu dutenen proportzioaren inguruko ustea balio handiago batean zentratuta dago orain.

2.5 Ondorengo banaketa

A posteriori banaketak datuak behatu ondoren parametroaz dugun ezagutza adierazten du, eta gure a priori ezagutza (aldez aurretiko banaketa) eta datuak (egiantza) biltzen ditu.

Normalean, ez da erraza izaten banaketa bat lehen begi-kolpean interpretatzea, eta horretan lagunduko diguten parametroak behar ditugu. Halaber, sinesgarritasun-tarte bayestarrak kalkulatzen dira; tarte bayestar horiek, konfiantza-tarte tradizionalek ez bezala, probabilitate kontzeptuarekin lan egiteko aukera ematen digute.

2.5.1 Posizio-neurria eta zentrorako joeraren neurria

Lehenik eta behin, hiru posizio-neurririk ezagunenak hartuko ditugu kontuan: ondorengo moda, ondorengo mediana eta ondorengo batezbestekoa.

A posteriori moda ondorengo banaketa maximizatzen duen balioa da. Banaketa jarraitua bada, maximoak dentsitate-banaketa deribatu eta lortutako adierazpena 0rekin berdintzen du. Ondorengo banaketa $beta(a', b')$ bat denean, deribatua honela adierazten da:

$$\mathrm{g'(\pi|y)} = (\mathrm{a'} - 1)\pi^{a'-2} * (1-\pi)^{b'-1} + \pi^{a'-1} * (-1)(b'-1)(1-\pi)^{b'-2}.$$

$\mathrm{g'(\pi|y)}$ 0rekin berdintzen badugu, honako hau lortuko dugu:

$$moda = \frac{a'-1}{a'+b'-2}.$$

Hala ere, modak alderdi negatibo batzuk ditu posizio-neurri gisa. Lehenik eta behin, moda banaketaren mutur batetik hurbil egon daiteke, eta kasu horretan ez litzateke oso adierazgarria izango. Gainera, banaketak hainbat maximo lokal izan ditzake, eta prozesu honekin maximo eta minimo lokal guztiak aurkituko genituzke.

Ondorengo mediana datu guztiak txikitik handira ordenatuta daudenean datu horien erdian dagoen balioa da. Baldin eta $\mathrm{g}(\pi|\mathrm{y})$ $beta(a', b')$ bat bada, mediana honako honen soluzioa da:

$$\int_0^{mediana} \mathrm{g}(\pi|\mathrm{y})d\pi = 0.5.$$

Mediana lokalizazio-neurri bikaina da, baina alde txar bat dauka: zenbakizko kalkuluak egin behar dira lortzeko.

Azkenik, batezbestekoa oso neurri erabilia da. Batezbestekoa a posteriori banaketan espero den balioa da.

$$m' = \int_0^1 \pi\, \mathrm{g}(\pi|\mathrm{y})d\pi. \qquad (2.4)$$

Batezbestekoa nabarmen aldatzen da banaketak ilara astun bat duenean. Are gehiago, banaketa alboratuta badago eta ilara astun bat badauka, batezbestekoaren balioa balio gehienetatik oso urrun gera daiteke. Baldin eta $\mathrm{g}(\pi|\mathrm{y})$ $beta(a', b')$ bat bada, batezbestekoak honako balio hau izango du:

$$m' = \frac{a'}{a' + b'}. \qquad (2.5)$$

$beta(a', b')$ banaketa 0ren eta 1en artean bornatuta dago, eta, beraz, ez du ilara astunik. Hortaz, batezbestekoa posizio-neurri egokia izango da a posteriori beta banaketetarako.

2.5.2 Sakabanatze-neurriak

A posteriori banaketa batean aztertu beharreko bigarren alderdi interesgarria banaketa horrek dituen balioen sakabanatzea da. Oso sakabanatuta badaude, parametroaren informazioa ez da zehatza izango, datuen analisian lortutakoa gorabehera.

Ondorengo bariantza honela adierazten da:

$$Var(\pi|\mathrm{y}) = \int_0^1 (\mathrm{a}' - 1)g(\pi|\mathrm{y})d\pi. \qquad (2.6)$$

Banaketa $beta(a', b')$ bat bada,

$$Var(\pi|\mathrm{y}) = \frac{a'b'}{(a' + b')^2(a' + b' + 1)}. \qquad (2.7)$$

Bariantza, batezbestekoa bezala, nabarmen aldatzen da ilara astunak dituzten banaketekin. Baldintza horietan, bariantza oso handia izango da, probabilitate gehienak banaketaren erdialdetik hurbil samar kontzentratuta badaude ere. Bestalde, bariantza unitate karratuetan ematen da, eta,

beraz, haren interpretazioa ez da oso intuitiboa. Arazo horiek konpontzeko, askoz ohikoagoa da ondorengo desbiderapen estandarrarekin (hots, bariantzaren erro karratuarekin) lan egitea.

A posteriori banaketaren k-garren pertzentila π_k balioa da, hau da, bere azpiko azaleraren % k biltzen duena. Balio hori honako adierazpen hau zenbakiz ebazteko erabiltzen da:

$$k = 100 * \int_{-\infty}^{\pi_k} g(\pi|y)d\pi.$$

Pertzentil batzuk bereziki garrantzitsuak dira: lehen kuartila (Q_1) 25. pertzentila da, bigarren kuartila edo mediana 50. pertzentila, eta hirugarren kuartila (Q_3) 75. pertzentila.

Kuartilarteko heina ($IQR = Q_3 - Q_1$,) ilara astunek aldatzen ez duten sakabanatze-neurri bat da.

2.4.6 Adibidea (jarraipena)

Aurreko adibidera itzuliz, posizio-neurri eta sakabanatze-neurri batzuk kalkulatuko ditugu. Zuzendariaren, ingeleseko irakaslearen eta amaren kasuan, batezbestekoa eta desbiderapena zuzenean kalkula ditzakegu (2.4) eta (2.6) formulak erabiliz. Matematikako irakaslearen kasuan, berriz, (2.3) eta (2.5) ekuazioak erabili behar ditugu, eta zenbakiz ebatzi. Ondorengo taulan, lortutako neurrietako batzuk agertzen dira. Ondoriozta dezakegunez, amaren kasurako zenbatetsitako balioak izan ezik, ondorengo balioak antzeko samarrak dira, a priori banaketa desberdinetatik abiatu badira ere. Azpimarratu beharra dago ingeleseko irakaslea eta ama aldez aurretiko batez besteko proportzio beretik abiatu direla, baina ondorengo batezbestekoak desberdin samarrak direla. Diferentzia hori amak aldez aurretiko informazio gehiago zeukalako gertatu da.

Pertsona	Banaketa	Batezbestekoa	Mediana	Desbid. est.	IQR
Zuzendaria	Beta(19.2, 4.8)	0.8	0.8084153	0.08	0.1087928
	Beta(93.2, 30.8)	0.7516129	0.7529691	0.03864618	0.05229502
Ingeleseko irakaslea	Beta(1, 1)	0.5	0.5	0.2886751	0.5
	Beta(75, 27)	0.7352941	0.7368369	0.04347041	0.05888509
Matem. irakaslea	Diskretua	0.66	0.6	0.02666667	0.2
	Zenbakizkoa	0.7297143	0.7297702	0.04255308	0.05694306
Ama	Beta(24.5, 24.5)	0.5	0.5	0.08838835	0.1216675
	Beta(98.5, 50.5)	0.6832061	0.684141	0.04049275	0.05484126

2.7 Sinesgarritasun-tarte bayestarrak

Sinesgarritasun-tarte bayestar bat zera da, parametro jakin bat edukitzeko $(1-\alpha)$ ondorengo probabilitatea duten balioen hein bat. Tarte horiek tradizionalen aldean duten abantaila nagusia probabilitatea zuzenean interpretatzeko aukera ematen dutela da.

2.7.1 π-rako sinesgarritasun-tarte bayestarra π

Dakigunez, $beta(a,b)$ banaketa batekin lan egiten bada, a posteriori banaketa $beta(a',b')$ bat izango da. Banaketa hori banaketa normal baten bidez hurbil daiteke, batezbestekoari eta desbiderapenari eutsiz.

$(\pi|y)$ da gutxi gorabehera $N(m',(s')^2)$.

Non

$$m' = \frac{a'}{a'+b'} \quad \text{y} \quad (s')^2 = \frac{a'b'}{(a'+b')^2(a'+b'+1)}.$$

Horrela, beraz, π-ren $(1-\alpha)*100\%$eko sinesgarritasun-eskualdea honako hau da gutxi gorabehera:

$$m' \pm z_{\frac{\alpha}{2}} * s'.$$

Non $z_{\frac{\alpha}{2}}$ banaketa normalaren taulatik lortzen dugun balioa den. 95%eko sinesgarritasun-tarte baterako, $z_{0.025} = 1.96$ da. Hurbilketa horrek oso portaera ona dauka denean. $a, b \geq 10$.

2.7.2 Adibidea (jarraipena)

Aurreko adibidearekin jarraituz, gainditu dutenen proportziorako % 95eko sinesgarritasun-tarteak kalkulatu ditugu. Lau kasuetan, formula zehatzaren (beta banaketa) eta hurbilketa normal baten bidez lortu ditugu. Taulako datuak behatu ondoren, hurbilketak balio zehatzetik oso hurbil daudela ondoriozta daiteke.

Pertsona	Ondorengo banaketa	Sinesgarritasun-tarte zehatza	Sinesgarritasun-tarte normala
Zuzendaria	Beta(93.2, 30.8)	(0.6857768, 0.8128175)	(0.6758678, 0.827358)
Ingeleseko irakaslea	Beta(75, 27)	(0.6612032, 0.804115)	(0.6500937, 0.8204946)
Matem. irak.	Zenbakizkoa	(0.6418581, 0.8086913)	(0.6463118, 0.8131168)
Ama	Beta(98.5, 50.5)	(0.6150072, 0.7482125)	(0.6038418, 0.7625704)

3. Batezbesteko normal baterako inferentzia bayestarra

3.1 Sarrera

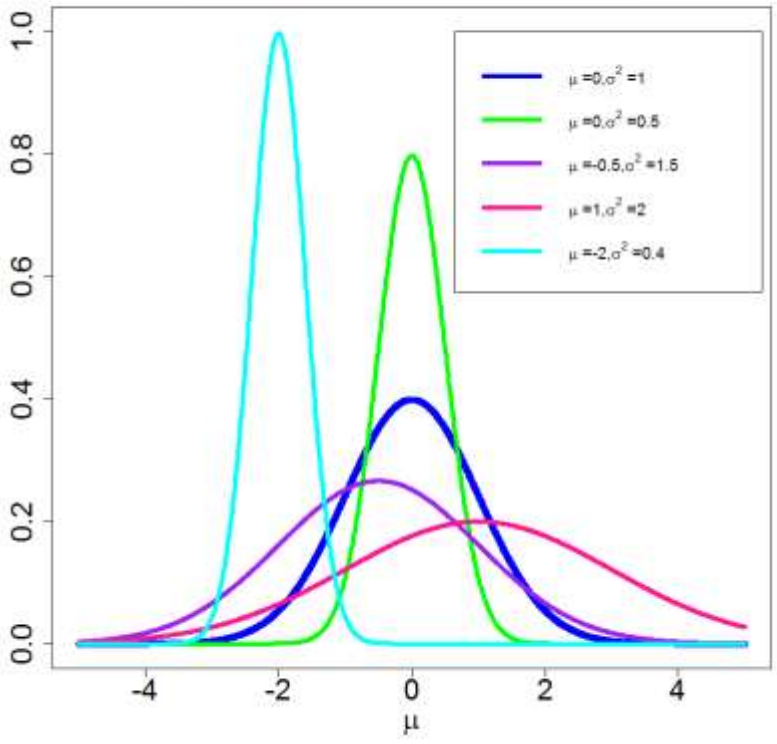
Estatistikan eta probabilitatean, fenomeno errealetan hurbilduta gehien agertzen diren aldagai jarraituko probabilitate-banaketetako bati deitzen zaio banaketa normal, Gaussen banaketa edo banaketa gausstarra. Batez ere horregatik garatu dira hainbeste metodo estatistiko banaketa mota horretarako.

$N(\mu, \sigma^2)$ banaketa baten batezbestekoa da, eta bariantza, berriz, σ^2; dentsitate-funtzioa honela adierazten da:

$$g(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \qquad \infty < x < \infty.$$

Banaketaren dentsitate-funtzioaren grafikoak kanpai-forma dauka, eta parametro estatistiko jakin batekiko simetrikoa da. Kurba horri Gaussen kanpai deitzen zaio, eta funtzio gausstar baten grafikoa da. Ondoren, hainbat batezbesteko eta bariantza dituzten kanpai batzuk ikusiko ditugu.

Banaketa Normala $N(\mu, \sigma^2)$



Probabilitateen banaketa normalen "familia" bat dago. Dakigunez, banaketa bakoitzak batezbesteko edo desbiderapen estandar desberdina izan dezake. Beraz, banaketa normalen kopurua mugagabea da, eta ezinezkoa izango litzateke μ eta σ-ren konbinazio bakoitzerako probabilitate-taula bat ematea. Arazo hori konpontzeko, banaketa normalen familiako "kide" bakarra erabiltzen da, hots, 0

batezbestekoa eta 1 desbiderapen estandarra duena; banaketa horri **N(0,1) banaketa estandar normal** deitzen zaio. Horrela, beraz, behaketa bakoitzeko batezbestekoa kenduz eta desbiderapen estandarrarekin zatituz gero, banaketa normal guztiak banaketa estandar bihur daitezke.

3.1.1 Mugaren Teorema Zentrala

Teorema honen arabera, μ batezbestekoa eta σ^2 bariantza dituen edozein banaketatik hartutako $y_1, \dots, y_n$ ausazko lagin batetik abiatuta, $\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$ banaketa $N(0,1)$ banaketa normal bat da mugan. Hau da, banaketak mugan duen forma banaketa normal estandar batena izango da, jatorrizko banaketa normala izan ala ez. Harrigarria bada ere, "muga" kontzeptua $n \geq 25$ den kasuan aplika daiteke.

3.2 Bayesen teorema aldez aurretiko diskretu bat duen batezbesteko normal baterako

3.2.1 Behaketa bakarrarekin

Har dezagun behaketa bakar bat $f(y|\mu)$ dentsitate-funtziotik, σ^2. bariantza ezaguna duen banaketa normal bat dela suposatuta. σ desbiderapen estandarra bariantzaren erro karratua da, eta m balio posible ditugu batezbestekorako: μ_1,..., μ_m. Balio horietatik abiatuta, behaketaren aurretik genuen ustea adieraziko duen aldez aurretiko banaketa bat aukeratzen da. Aldez aurretiko informaziorik izan ezean, arrakasta-probabilitate bera emango diegu balio guztiei.

Egiantzak pisuak ematen dizkie balio posible guztiei, gertatzeko probabilitatearen arabera. Horretarako, behatutako balioa finkatu eta balio posible guztien arteko aldagaia aldatu behar dugu. Ondorengo banaketa aldez aurretiko banaketaren eta egiantzaren biderkaduraren proportzionala da.

$$g(\mu|y) = \frac{prior * likelihood}{\sum prior * likelihood}$$

3.2.1.1 Behaketa bakarraren egiantza

$y|\mu$-ren baldintzazko banaketa normala da, eta batezbestekoa eta σ^2 bariantza ezaguna ditu. Gainera, dentsitate-funtzioa honako hau da:

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Egiantzaren forma honela adierazten da:

$$f(y|\mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \qquad (3.1)$$

Adierazpen horretan, *y*-k konstante jarraitzen du behatutako balioarekin, eta μ balio posibleen artean dago. Egiantzaren balioa lortzeko, bi aukera ditugu. Alde batetik, banaketa erreala banaketa normal estandar bihur daiteke Z balioa edo Z estatistikoa erabiliz, hau da, hautatutako balioaren (X) eta μ batezbestekoaren arteko distantzia, σ desbiderapen estandarrarekin zatitua.

Formalki, baldin eta $X \sim N(\mu, \sigma^2)$ bada, orduan ausazko aldagaia $z = \frac{X-\mu}{\sigma}$ 0 batezbestekoa eta 1 desbiderapen estandarra duen banaketa normal estandar edo tipifikatu baten arabera banatzen da: $N(0,1)$. Adierazi beharra dago normal estandarraren banaketa 0ren simetrikoa denez gero, $f(z) = f(-z)$. dela. Bigarren aukera, bestalde, egiantza zuzenean (3.1) formularen bidez lortzean datza.

3.2.1.2 Adibidea

Demagun batezbestekoa eta $\sigma^2 = 1$ bariantza ezaguna dituen $y|\mu$ normala dela eta -k 2, 2.5, 3, 3.5 eta 4 balioak probabilitate berdinarekin har ditzakeela. Behaketa bat ausaz hartu, eta $y = 3.2$.

Lehenik eta behin, $z = \frac{y-\mu}{\sigma}$ dela jakinik eta normalaren ohiko taularako emandako datuak kontuan hartuta, honako taula hau sor dezakegu:

	Aldez aurretikoa	Z	Egiantza	Aldez aurretikoa*Egiantza	Ondorengoa
2	0.2	-1.2	0.1942	0.03884	0.1238
2.5	0.2	-0.7	0.3123	0.06246	0.1991
3	0.2	-0.2	0.3910	0.0782	0.2493
3.5	0.2	0.3	0.3814	0.07628	0.2431
4	0.2	0.8	0.2897	0.05794	0.1847
				0.31372	1

(3.1) ekuazioaren bidez, berriz, honako taula hau dugu (biribiltze-erroreak salbuetsita, baliokideak dira):

	Aldez aurretikoa	Egiantza	Aldez aurretikoa*Egiantza	Ondorengoa
2	0.2	$e^{-\frac{1}{2}(3.2-2)^2} = 0.4868$	0.0974	0.1239
2.5	0.2	$e^{-\frac{1}{2}(3.2-2.5)^2} = 0.4868$	0.1565	0.1990
3	0.2	$e^{-\frac{1}{2}(3.2-3)^2} = 0.9802$	0.1960	0.2493
3.5	0.2	$e^{-\frac{1}{2}(3.2-3.5)^2} = 0.9560$	0.1912	0.2432
4	0.2	$e^{-\frac{1}{2}(3.2-4)^2} = 0.7261$	0.1452	0.1846
			0.7863	1

3.2.2 Ausazko lagin batekin

Normalean, behaketa bakar baten ordez $y_1, y_2, \ldots, y_n$ behaketa izaten ditugu ausazko lagin batetik abiatuta, non $y_j|\mu$ banaketa bakoitza normala den, batezbestekoarekin eta σ^2. bariantza ezagunarekin. Ondorengo banaketa aldez aurretikoaren eta egiantzaren arteko biderkaduraren proportzionala da, eta laginaren behaketak independenteak direnez gero, baterako egiantza behaketa guztien egiantzaren biderkadurak ematen du:

$$f(y_1, \ldots, y_n|\mu) = f(y_1|\mu) * f(y_2|\mu) *\ldots* f(y_n|\mu).$$

Kasu honetan, beraz, honela adierazten da Bayesen teorema, aldez aurretiko diskretu batekin:

$$g(\mu|y_1, \ldots, y_n) \propto g(\mu) * f(y_1|\mu) * f(y_2|\mu) *\ldots* f(y_n|\mu).$$

3.2.2.1 Ondorengo probabilitateak

Ondorengo probabilitateak banan banan bila daitezke, era sekuentzialean edo denak batera. Lehen kasuan, behaketa baten ondorengoa hurrengo behaketaren aldez aurretiko bihurtzen da, baina metodo hori oso neketsua izan daiteke laginaren tamaina handitzen den heinean. Bigarren kasuan, behaketa bakoitza normala denez gero, guztiek egiantz normala dute, eta baterako egiantza honela adierazten da:

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{1}{2\sigma^2}(y_1-\mu)^2} * e^{-\frac{1}{2\sigma^2}(y_2-\mu)^2} *\ldots* e^{-\frac{1}{2\sigma^2}(y_n-\mu)^2},$$

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{1}{2\sigma^2}[(y_1-\mu)^2+(y_2-\mu)^2+\cdots+(y_n-\mu)^2]}$$

Parentesi artean dauden gaiei dagokienez:

$$[(y_1-\mu)^2 + (y_2-\mu)^2 + \cdots + (y_n-\mu)^2] = {y_1}^2 - 2y_1\mu + \mu^2 + \cdots + {y_n}^2 - 2y_n\mu + \mu^2 =$$

$$= ({y_1}^2 + \cdots + {y_n}^2) - 2\mu(y_1 + \cdots + y_n) + n\mu^2 = n\left[\frac{({y_1}^2 + \cdots + {y_n}^2)}{n} - 2\mu\bar{y} + \mu^2\right]$$

Orduan,

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{n}{2\sigma^2}\left[\mu^2-2\mu\bar{y}+\bar{y}^2-\bar{y}^2+\frac{({y_1}^2+\cdots+{y_n}^2)}{n}\right]},$$

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{n}{2\sigma^2}[\mu^2-2\mu\bar{y}+\bar{y}^2]} e^{-\frac{n}{2\sigma^2}\left[\frac{({y_1}^2+\cdots+{y_n}^2)}{n}-\bar{y}^2\right]}.$$

$y_1, y_2, \ldots, y_n$ ausazko lagin normal baten egiantza $\bar{y}$ lagin-batezbestekoaren egiantzaren proportzionala da. parametroaren mende ez dagoen zatia sinplifikatuz,

$$f(y_1, \ldots, y_n|\mu) \propto e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2}.$$

Ikusten denez, batezbestekoa eta $\frac{\sigma^2}{n}$ bariantza dituen normal baten forma dauka egiantza horrek. $\bar{y}$ lagin-batezbestekoa ere batezbestekoa eta $\frac{\sigma^2}{n}$ bariantza dituen normal baten bidez banatuta dago, eta, beraz, ausazko laginaren baterako egiantza lagin-batezbestekoaren egiantzaren proportzionala da, honela adierazita:

$$f(\bar{y}|\mu) \propto e^{-\frac{1}{\frac{2\sigma^2}{n}}(\bar{y}-\mu)^2} . \qquad (3.2)$$

Hurrengo kapituluan ikusiko dugunez, formula horren bidez errazagoa eta eraginkorragoa da a posteriori probabilitateak kalkulatzea.

3.2.2.2 Adibidea (jarraipena)

Demagun orain 4ko tamainako ausazko lagin bat hartu dugula batezbestekoa eta σ^2 bariantza dituen banaketa normal batetik: 3.2, 2.2, 3.6 eta 4.1. Lagin-batezbestekoa $\bar{y} = 3.275$ da, eta, haren egiantza erabiliz, honako emaitza hauek lor ditzakegu:

	Aldez aurretikoa	Egiantza	Aldez aurretikoa*Egiantza	Ondorengoa
2	0.2	$e^{-\frac{1}{2*1/4}(3.275-2)^2} = 0.0387$	0.0077	0.0157
2.5	0.2	$e^{-\frac{1}{2*1/4}(3.275-2.5)^2} = 0.3008$	0.0602	0.1228
3	0.2	$e^{-\frac{1}{2*1/4}(3.275-3)^2} = 0.8596$	0.1719	0.3505
3.5	0.2	$e^{-\frac{1}{2*1/4}(3.275-3.5)^2} = 0.9037$	0.1807	0.3685
4	0.2	$e^{-\frac{1}{2*1/4}(3.275-4)^2} = 0.3495$	0.0699	0.1425
			0.4904	1

3.3 Bayesen teorema aldez aurretiko jarraitu bat duen batezbesteko normal baterako

Aurreko atalean, -k har zitzakeen balioak balio kopuru jakin batera mugatuta zeuden, baina errealistagoa da pentsatzea -k edozein balio har dezakeela edo, gutxienez, tarte jakin batean dauden guztiak. Kasu horietan, komeni da aldez aurretiko banaketa jarraitu bat erabiltzea. Honela idatz dezakegu Bayesen teorema:

$$g(\mu|y_1,\ldots,y_n) \propto g(\mu) * f(y_1,\ldots,y_n|\mu)$$

Aldez aurretikoa diskretua zenean, egiantzaren eta aldez aurretikoaren biderkadura -ren balio guztien biderkaduren baturarekin zatitu behar genuen ondorengoa lortzeko. Banaketa jarraitua denean, prozedura baliokidea biderkadura horren integrala egitea da. Hau da:

$$g(\mu|y_1,\ldots,y_n) = \frac{g(\mu) * f(y_1,\ldots,y_n|\mu)}{\int g(\mu) * f(y_1,\ldots,y_n|\mu)d\mu} \qquad (3.3)$$

Banaketa normal baterako, ausazko laginaren egiantza $\bar{y}$ lagin-batezbestekoaren egiantzaren proportzionala da. Orduan,

$$g(\mu|y_1,\ldots,y_n) = \frac{g(\mu) * e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2}}{\int g(\mu) * e^{-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2} d\mu}.$$

3.3.1 -rako aldez aurretiko dentsitate laua (Jeffrey-ren aldez aurretikoa)

Aldez aurretiko lauan, -k har ditzakeen balio guztiak probabilitate berekoak dira, eta guztiei pisu bera ematen zaie: $g(\mu) = 1$. Aldez aurretiko laua ez da egiaz aldez aurretiko banaketa bat, eremuaren integrala ez baita 1, baina hori ez da arazoa. Gainera, a posteriori banaketaren integrala 1 izango da, eta hori da interesatzen zaiguna.

3.3.1.1 Behaketa normal bat

Demagun y behaketa batezbestekoa eta σ^2 bariantza dituen normal bat bezala banatutako behaketa bat dela. Proportzionaltasun-konstantea alde batera utzita, egiantza honela lortzen da:

$$f(y|\mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2}.$$

Aldez aurretikoa beti 1 denez gero, ondorengoa aurreko adierazpenaren proportzionala da.

$$g(\mu|y) \propto e^{-\frac{1}{2\sigma^2}(\mu-y)^2}.$$

Ekuazio horretatik abiatuta, ondoriozta dezakegu ondorengoak y batezbestekoa eta σ^2 bariantza dituen banaketa normal bati jarraitzen diola.

3.3.1.2 Ausazko lagin normal bat

Aurreko atal batean ikusi dugunez, banaketa normal batetik hartutako ausazko lagin baten egiantza $\bar{y}$ lagin-batezbestekoaren egiantzaren proportzionala da. Dakigunez, $\bar{y}$ normal banatuta dago, batezbestekoarekin eta $\frac{\sigma^2}{n}$ bariantzarekin. Orduan, konstantea alde batera uzten badugu, egiantza honela lortzen da:

$$f(\bar{y}\,|\mu) \propto e^{-\frac{1}{2\sigma^2/n}(\bar{y}\,-\mu)^2}.$$

Aldez aurretikoa beti 1 denez gero, ondorengoa aurreko adierazpenaren proportzionala da.

$$g(\mu\,|\bar{y}) \propto e^{-\frac{1}{2\sigma^2/n}(\mu-\bar{y}\,)^2}$$

Ekuazio horretatik abiatuta, ondoriozta dezakegu ondorengoak $\bar{y}$ batezbestekoa eta $\frac{\sigma^2}{n}$ bariantza dituen banaketa normal bati jarraitzen diola.

3.3.2 -rako aldez aurretiko dentsitate normala

3.3.2.1 Behaketa normal bat

Demagun batezbestekoa eta σ^2 bariantza ezaguna dituen banaketa normal batetik hartutako y behaketa bat dugula. Demagun, halaber, m batezbestekoa eta s^2 bariantza dituen aldez aurretiko normal bat dugula. ez duen zatia alde batera utzita, aldez aurretikoa eta egiantza honela lortzen dira:

$$g(\mu) \propto e^{-\frac{1}{2s^2}(\mu-m)^2},$$

$$f(y\,|\mu) \propto e^{-\frac{1}{2\sigma^2}(y-\mu)^2}.$$

Bi ekuazioen biderkadura honela adierazten da:

$$g(\mu) * f(y\,|\mu) \propto e^{-\frac{1}{2}\left[\frac{(\mu-m)^2}{s^2}+\frac{(y-\mu)^2}{\sigma^2}\right]}$$

$$\propto e^{-\frac{1}{2}\left[\frac{\sigma^2(\mu^2-2\mu m+m^2)+s^2(y^2-2y\mu+\mu^2)}{s^2\sigma^2}\right]}$$

$$\propto e^{-\frac{1}{2}\left[\frac{(\sigma^2+s^2)\mu^2-2(\sigma^2 m+s^2 y)\mu+m^2\sigma^2+y^2s^2)}{s^2\sigma^2}\right]}$$

$\frac{(\sigma^2+s^2)}{\sigma^2 s^2}$ gaia faktore komun gisa ateratzen, formula nabarmenak erabiltzen eta -ren mende ez dauden gaiak xurgatzen baditugu, honako hau lortuko dugu:

$$\propto e^{-\frac{1}{2\sigma^2s^2/(\sigma^2+s^2)}\left[\mu^2-2\frac{(\sigma^2m+s^2y)}{s^2\sigma^2}\mu+\left(\frac{(\sigma^2m+s^2y)}{\sigma^2+s^2}\right)^2\right]}$$

$$\propto e^{-\frac{1}{2\sigma^2s^2/(\sigma^2+s^2)}\left[\mu^2-\frac{(\sigma^2m+s^2y)}{\sigma^2+s^2}\right]}$$

Horretan oinarrituta, ondoriozta dezakegu ondorengo banaketak honako batezbesteko eta bariantza hauek dituela:

$$m' = \frac{(\sigma^2 m + s^2 y)}{\sigma^2 + s^2}, \qquad (s')^2 = \frac{\sigma^2 s^2}{(\sigma^2 + s^2)}. \qquad (3.4)$$

Hau da, $N(m, s^2)$ aldez aurretikoarekin hasi eta $N(m', (s')^2)$ ondorengoarekin amaitu dugu. Horrek esan nahi du Bayesen teorema aplikatzean ez dela beharrezkoa izango integrala egitea, familia bereko bi banaketa baitira.

3.3.1.3 Normal baten balioak eguneratzea

(3.4) ekuazioko adierazpenak sinplifikatu egin daitezke. Aurrenik, banaketaren zehaztasuna definituko dugu, hau da, bariantzaren alderantzizko balioa. Zehaztasunak batukortasun-propietatea betetzen du, eta ondorengoa honela adierazten da:

$$\frac{1}{(s')^2} = \left(\frac{\sigma^2 s^2}{(\sigma^2 + s^2)}\right)^{-1} = \frac{(\sigma^2 + s^2)}{\sigma^2 s^2} = \frac{1}{s^2} + \frac{1}{\sigma^2}$$

Hau da, ondorengo zehaztasuna aldez aurretiko zehaztasunaren eta behaketaren zehaztasunaren arteko baturaren berdina da. Ondorengo batezbestekoa, berriz, honela adierazten da:

$$m' = \frac{(\sigma^2 m + s^2 y)}{\sigma^2 + s^2} = \frac{\sigma^2}{\sigma^2 + s^2} * m + \frac{s^2}{\sigma^2 + s^2} * y.$$

Sinplifikazioren bat eginez, ikusiko dugu ondorengo batezbestekoa aldez aurretiko batezbestekoaren eta behaketaren batezbestekoaren batezbesteko haztatua dela, pisuak ondorengo zehaztasunak izanik.

$$m' = \frac{1/s^2}{1/\sigma^2 + 1/s^2} * m + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{1^2}{s}} * y.$$

Aldez aurretiko lauaren kasuan, bariantza infinitua da, eta, beraz, zehaztasuna 0 da. Hortaz, ondorengo zehaztasuna eta aldez aurretiko zehaztasuna berdinak izango dira, eta a posteriori bariantzak ere bai (σ^2).

$$\frac{1}{\sigma^2} = 0 + \frac{1}{\sigma^2}$$

Hala ere, aldez aurretiko horren a priori batezbestekoa ez dago oso ongi zehaztuta. Beraz, a posteriori batezbestekoak behaketaren balio bera duela ulertuko da. y.

3.3.1.2 Ausazko lagin normal bat

Demagun batezbestekoa eta σ^2 bariantza ezaguna dituen banaketa normal batetik hartutako $y_1, y_2, \ldots, y_n$ ausazko lagin bat dugula. Honako aldez aurretiko banaketa normal hau izango dugu, m batezbestekoarekin eta s^2 bariantzarekin:

$$g(\mu) \propto e^{-\frac{1}{2s^2}(\mu - m)^2}.$$

$\bar{y}$ lagin-batezbestekoaren egiantza erabiliko dugu; batezbesteko hori batezbestekoa eta σ^2/n bariantza ezaguna dituen normal bat bezala banatuta dago. $\bar{y}$-ren zehaztasuna $\frac{n}{\sigma^2}$ da, hots, behaketen zehaztasun guztien batura. Gainera, ondorengo zehaztasuna aldez aurretiko zehaztasunaren eta $\bar{y}$-ren zehaztasunaren arteko baturaren berdina da, eta ondorengo bariantza ondorengo zehaztasunaren erreziprokoa da.

$$\frac{1}{(s')^2} = \frac{1}{s^2} + \frac{n}{\sigma^2} = \frac{(\sigma^2 + ns^2)}{\sigma^2 s^2} \qquad (3.5)$$

Azkenik, ondorengo batezbestekoa aldez aurretiko batezbestekoaren eta $\bar{y}$-ren batezbestekoaren batezbesteko haztatua da, pisuak ondorengo zehaztasunak izanik:

$$m' = \frac{1/s^2}{n/\sigma^2 + 1/s^2} * m + \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} * \bar{y}. \qquad (3.6)$$

3.4 Aldez aurretiko normala aukeratzea

Aldez aurretiko banaketa aukeratzeko garaian, gure ezagutza ahal bezain ongien adieraztea da helburua. Bariantza ezaguneko banaketa normal batetik hartutako behaketen kasuan, -rako banaketak $N(m, s^2)$ dira. Gure usteak horrelako banaketa baten bidez adierazten baditugu, erraza izango da gure ezagutza Bayesen teoremarekin eta (3.5) eta (3.6) ekuazioekin eguneratzea, ez baita beharrezkoa izango integralak zenbakiz kalkulatzea.

Lehen urratsa gure ustea, m aldez aurretiko batezbestekoa eta s aldez aurretiko desbiderapen estandarra zer baliotan zentratuta dauden erabakitzea izango da. Horretarako, gure ezagutzaren arabera balio minimo eta maximo posibleak hartzea *(m+x, m-x)*, haien diferentzia ateratzea eta emaitza 6rekin zatitzea izango da errazena. Lortutako balioak arrazoizko probabilitate bat emango die posibletzat hartu ditugun balioei. Hau da, $sd = \frac{(m+x)-(m-x)}{6} = \frac{2x}{6} = \frac{x}{3}$ izango da. Prozedura horren helburua $P(m < m + x) = 0.4987$ probabilitatea bilatzea eta honako adierazpen hau tipifikatzea da: $P\left(Z < \frac{x}{s}\right) = 0.4987$. Normal estandarraren probabilitateen taulan oinarrituta, $\frac{x}{s} = 3 \Longrightarrow s = \frac{x}{3}$ lortuko dugu.

Berriz ere, komeni da aldez aurretiko banaketaren lagin-tamaina baliokidea kontuan hartzea, behar bezala egokitzen den jakiteko. $n_{eq} = \frac{\sigma^2}{s^2}$. n_{eq} handia izateak esan nahi du gure ezagutzak esanguratsuak izango direla, eta ebidentzia berri ugari beharko dira balio horietatik urruntzeko. n_{eq} txikia bada, berriz, ondorengo ezagutzak behatutakoaren eragin handia izango du.

Gure ezagutza behar bezala hurbilduko duen banaketa normalik aurkitu ezean, balio batzuk finkatu (bilatzen dugun balioa balio horien artean egotea espero dugu) eta elkarren artean interpolatu daitezke aldez aurretikoa lortzeko. Kasu horietan, ondorengo banaketa honako hau izango da:

$$g(\mu|y_1, y_2, \ldots, y_n) = \frac{f(y_1, y_2, \ldots, y_n|\mu) * g(\mu)}{\int f(y_1, y_2, \ldots, y_n|\mu) * g(\mu) d\mu}.$$

3.4.1 Adibidea

Silvia, Leire, Ane eta Sara urmael baten aurrean daude, eta bertan dauden karpa gorrien luzeraren batezbestekoa kalkulatu nahi dute. Aldez aurretiko azterlan batzuengatik dakigunez, 2 cm-ko desbiderapen estandar ezaguna duen normal baten bidez banatzen da luzera. Silviak 40 cm-tan finkatu du bere aldez aurretiko batezbestekoa, eta uste du ezinezkoa dela luzera $[28,52]$. tartetik kanpo egotea. Hortaz, $s = \frac{52-28}{6} = \frac{24}{6} = 4$ izango da, eta, beraz, Silviak $N(40, 4^2)$. bat erabiliko du. Leirek, berriz, ez daki ezer karpei buruz, eta aldez aurretiko lau bat erabiltzea erabaki du. Anek, bestalde, bere aldez aurretikoak banaketa normala eduki ordez forma trapezoidala edukiko duela erabaki du. Horrela, beraz, 0ko pisua eman dio 8 balioari, 1eko pisua 14-30 balioei, eta berriz ere 0koa 36 balioari. Azkenik, Sarak gogoratu du behin 50 cm-ko karpa bat ikusi zuela, eta informazio horretan oinarritu du bere ustea. Laginaren tamaina baliokidea 1 denez gero, bariantza honako hau izango da: $s^2 = \frac{\sigma^2}{n_{eq}} = 2^2$. Hortaz, aukeratutako banaketa $N(50, 2^2)$ bat izango da.

Ondoren, 12 karpako ausazko lagin bat hartu eta laginaren batezbestekoa kalkulatu dute: $\bar{y} = 42$ cm. Silviak, Leirek eta Sarak beren batezbestekoa eta ondorengo desbiderapena kalkulatu dute (3.5) eta (3.6) ekuazioen bidez. Lehen kasuan,

$$\frac{1}{(s')^2} = \frac{1}{4^2} + \frac{12}{2^2} \Longrightarrow s' = 0.5714,$$

$$m' = \frac{\frac{1}{4^2}}{\frac{1}{0.5714^2}} * 20 + \frac{\frac{12}{2^2}}{\frac{1}{0.5714^2}} * 22 = 41.95699$$

Aldez aurretiko lauaren kasuan:

$$\frac{1}{(s')^2} = \frac{12}{2^2} \Longrightarrow s' = 0.5774, \quad m' = 42.$$

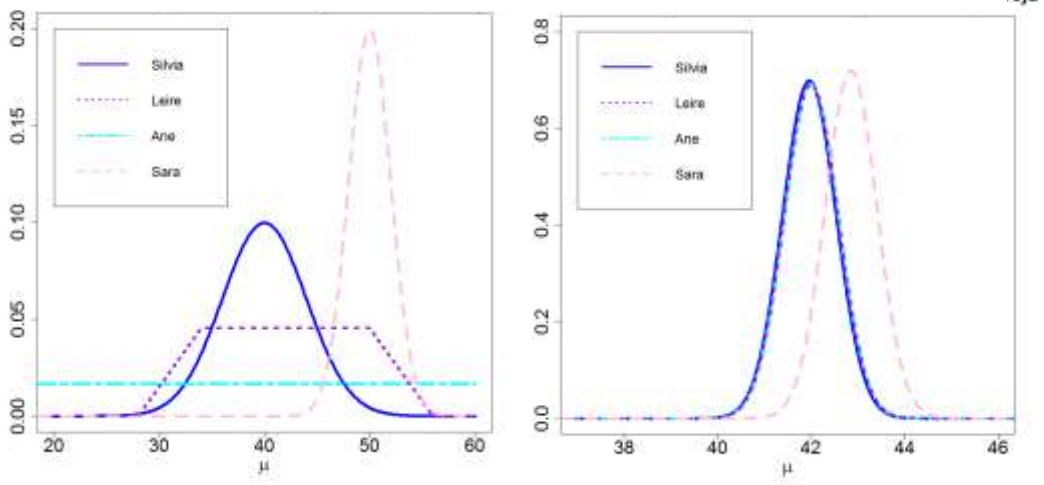
Eta Sararen kasuan:

$$\frac{1}{(s')^2} = \frac{1}{2^2} + \frac{12}{2^2} \Longrightarrow s' = 0.5547,$$

$$m' = \frac{\frac{1}{2^2}}{\frac{1}{0.5547^2}} * 40 + \frac{\frac{12}{2^2}}{\frac{1}{0.5547^2}} * 42 = 42.85086$$

Hiru a posteriori banaketak normalak dira. Dena dela, Aneren kasuan (3.3) ekuazioa erabili behar da, eta, horretarako, zenbakizko integrazioa behar da. Ondorengo grafikoetan, gure datuen aldez aurretiko eta ondorengo banaketak ditugu irudikatuta. Azpimarratu beharra dago aldez aurretikoak marrazteko garaian Leire eta Aneren kurben azpiko eremua bat izatea eragiten duen konstantea hartu dugula kontuan.

Karpa gorriaren luzeraren aldez aurretiko banaketak **Karpa gorriaren luzeraren ondorengo banaketak** roja



Lehen hiru banaketak ia berdinak dira, aldez aurretiko desberdinetatik abiatu badira ere. Hain zuzen ere, banaketak balio beraren inguruan zentratuta zeudelako eta aldez aurretiko ezagutza zehaztugabeetan oinarrituta zeudenez gero ezagutza horiek behatutako emaitzen eragin handia jaso dutelako gertatu da hori. Sararen ustea, berriz, alboratuago zegoen, baina behatutako 12 karpen emaitzen ondorioz, haren ezagutzak balio nabarmen txikiago batean zentratuta daude orain.

3.5 Batezbesteko normal baterako sinesgarritasun-tarte bayestarrak

Zenbaitetan, interesgarria izan daiteke gure "post-data" ezagutza balio-hein jakin batean laburbiltzea eta, horren arabera, probabilitate-maila batekin balioak barruan egotea. Kontzeptu horri tarte bayestar deitzen zaio. Normalean, ilara berdinak dituzten tarteekin lan egingo dugu.

3.5.1 Bariantza ezaguna

$y_1, y_2, \ldots, y_n$ behaketak $N(\mu, \sigma^2)$ banaketa batetik hartutako ausazko lagin bat direnean, $\bar{y}$ lagin-batezbestekoa $N(\mu, \sigma^2/n)$ da. Aldez aurretiko lau bat edo aldez aurretiko normal bat –$N(m, s^2)$– erabiliz, -ren ondorengo banaketa, $\bar{y}$ emanik, $(m', (s')^2)$ da.

-rako $(1-\alpha) * 100\%$-eko sinesgarritasun-tarte bayestarra honako hau da:

$$m' \pm z_{\frac{\alpha}{2}} * s'. \qquad (3.7)$$

Non $z_{\frac{\alpha}{2}}$ normalaren probabilitate-taulan aurkituko dugun. Hortaz, egiazko batezbestekoa tartearen barruan ez egoteko probabilitatea α da.

3.5.2 Bariantza ezezaguna

Bariantza ezezaguna denean, zehaztasuna ere ezezaguna da, eta, beraz, eguneratze-arauak ezin ditugu zuzenean aplikatu. Hasteko, laginaren bariantza kalkulatuko dugu:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Ondoren, (3.4) eta (4.5) ekuazioak erabiliko ditugu $(\mathrm{s}')^2$ eta m' aurkitzeko, eta $\widehat{\sigma}^2$ parametroa erabiliko dugu -ren ordez. σ^2.

Kasu horretan, ziurgabetasuna handiagoa da, σ^2 parametroa ere kalkulatu egin delako. Hori dela-eta, tarteek zabalagoak izan beharko lukete. Helburu horrekin, normalaren taula estandarra erabili ordez Studenten t banaketa erabiliko dugu, $df = n - 1$ askatasun-gradurekin. Hortaz:

$$m' \pm t_{\frac{\alpha}{2}} * s'. \qquad (3.8)$$

3.5.3 Aldez aurretiko ez-normala

Aldez aurretikoa normala ez denean, ondorengo banaketa ez-normal bat lortzen da Bayesen teoremaren eta zenbakizko integrazioaren bidez. Kasu horretan, $(1-\alpha) * 100\%$-eko sinesgarritasun-tartea lor daiteke, non μ_l-k eta μ_u-k honako hau beteko duten:

$$\int_{\mu_l}^{\mu_u} g(\mu|y_1, y_2, \ldots, y_n)d\mu = 1 - \alpha.$$

μ_l eta μ_u balioak ez dira bakarrak, eta, beraz, tarterik txikiena emango digutenak bilatuko ditugu. Gainera, honako hau ere bete beharko dute:

$$g(\mu_l|y_1, y_2, \ldots, y_n) = g(\mu_u|y_1, y_2, \ldots, y_n).$$

3.5.4 Adibidea (jarraipena)

Silviak, Leirek eta Sarak (3.7) ekuazioaren bidez kalkula dezakete 95%-eko sinesgarritasun-tartea. Anek, berriz, zenbakiz kalkulatu behar du. Ikus daitekeenez, tarteak antzekoak dira, lehen hiru kasuetan batez ere. Hortaz, behatutako datuen eragina aldez aurretikoaren eragina baino askoz handiagoa dela ondorioztatuko dugu.

Pertsona	Ondorengo banaketa	Sinesgarritasun-tartea
Silvia	$N(41.96, 0.5714^2)$	(41.01927- 42.8991)
Leire	$N(42, 0.5774^2)$	(41.05026- 42.94974)
Ane	Zenbakizkoa	(40.85819-43.11781)
Sara	$N(42.85, 0.5774^2)$	(41.93846-43.76326)

4. Batezbestekoen diferentziarako inferentzia bayestarra

Konparazioak funtsezkoak dira zientzia esperimentalean. Ausazko bi lagin konparatzeko, ez da eraginkorra lagin bakoitzean lortutako behaketak banan-banan konparatzea. Horregatik, lagin-batezbestekoak erabiliko ditugu konparazioak egiteko. Kasu gehienetan, banaketak normalak dira edo banaketa normal batera hurbil daitezke. Horregatik, bi banaketa normalen batezbestekoak konparatuko ditugu. Bestalde, lortutako ausazko laginak elkarren independenteak izaten dira maiz, eta, horregatik, independentziaren hipotesia betetzen dela suposatuko dugu kapitulu honetan.

4.1 Bariantza ezaguna eta berdina duten bi banaketa normalen ausazko lagin independenteak

Demagun σ^2 bariantza ezaguna dela. Bi laginak independenteak direnez gero, aldez aurretiko bi banaketa desberdin erabiliko ditugu batezbesteko bakoitzerako: $N(m_1, {s_1}^2)$ eta $N(m_2, {s_2}^2)$. Hortaz, ondorengo banaketak ere independenteak izango dira eta honela adieraziko dira:

$$\mu_1|y_{11}, \dots, y_{n_1 1} \sim N({m'}_1, ({s'}_1)^2)$$

$$\mu_2|y_{12}, \dots, y_{n_2 2} \sim N({m'}_2, ({s'}_2)^2)$$

Adierazpen horretan, ${m'}_i$ eta $({s'}_i)^2$ (4.5) eta (4.6) formuletatik lortzen dira ($i = 1{,}2$ izanik).

$\mu_1|y_{11}, \dots, y_{n_1 1} \sim N({m'}_1, ({s'}_1)^2)$ eta $\mu_2|y_{12}, \dots, y_{n_2 2} \sim N({m'}_2, ({s'}_2)^2)$ elkarren independenteak direnez gero, batezbestekoak eta bariantzak ausazko aldagai independenteetarako dituzten propietateak erabil daitezke. Horrek ondorengo banaketa hau ematen digu: $\mu_d = \mu_1 - \mu_2$. Hau da,

$$\mu_d|y_{11}, \dots, y_{n_1 1} y_{12}, \dots, y_{n_2 2} \sim N({m'}_d, ({s'}_d)^2)$$

Non ${m'}_d = {m'}_1 - {m'}_2$ eta $({s'}_d)^2 = ({s'}_1)^2 + ({s'}_2)^2$ diren. Hain zuzen ere, banaketa hori erabiliko dugu $\mu_1 - \mu_2$ batezbestekoen diferentzian inferitzeko.

4.1.1 Batezbestekoen diferentziarako sinesgarritasun-tarteak, bariantza ezaguna eta berdina izanik

Bariantza ezaguna denean, $z_{\frac{\alpha}{2}}$ balio kritikoa taula normal estandarizatutik lortzen da. Kasu horretan, honela adierazten da $(1-\alpha) * 100\%$-eko sinesgarritasun-tarte bayestarra $\mu_d = \mu_1 - \mu_2$-rako:

$$ {m'}_d \pm z_{\frac{\alpha}{2}} * {s'}_d. \qquad (4.1)$$

Eta honela berridatz daiteke:

$$m'_1 - m'_2 \pm z_{\frac{\alpha}{2}} * \sqrt{(s'_1)^2 + (s'_2)^2} \ . \qquad (4.2)$$

Hau da, $\mu_1 - \mu_2$ konfiantza-tartean egoteko probabilitatea $(1-\alpha) * 100\%$ da.

4.1.2 Aldebakarreko hipotesi-test bayestarra

Gure helburua μ_1 batezbestekoa μ_2 baino handiagoa den zehaztea bada, honako hipotesi-kontraste hau planteatu behar dugu:

$$\begin{cases} H_0\colon \mu_d \leq 0 \\ H_1\colon \mu_d > 0 \end{cases}$$

Non $\mu_d = \mu_1 - \mu_2$ batezbestekoen arteko diferentzia den. Era bayestarrean inferitzeko, hipotesi nuluaren ondorengo probabilitatea kalkulatu behar da, hots, $P(\mu_d \leq 0|datos)$, non $datos$ bi laginetako behaketak biltzen dituzten: $y_{11}, \dots, y_{n_1 1}$ eta $y_{12}, \dots, y_{n_2 2}$. Probabilitatea tipifikatuz, honako hau lortuko dugu:

$$P(\mu_d \leq 0|datos) = P\left(\frac{\mu_d - m'_d}{s'_d} \leq \frac{0 - m'_d}{s'_d}\right) = P\left(Z \leq \frac{0 - m'_d}{s'_d}\right). \qquad (4.3)$$

Adierazpen horretan, Z-k banaketa normal estandarrari jarraitzen dio. Normalaren probabilitate-taulatik abiatuta, $P\left(Z \leq \frac{-m'_d}{s'_d}\right)$ probabilitatea α baino txikiagoa bada, hipotesi nulua baztertu egin daiteke maila horretan. Kasu horretan, μ_1 batezbestekoa μ_2 baino handiagoa dela ondoriozta daiteke.

4.1.2 Aldebakarreko hipotesi-test bayestarra

Inferentzia bayestarrak ez du aukerarik ematen ondorengo probabilitatea kalkulatzearen bidez $H_0\colon \mu_d = 0$ versus $H_1\colon \mu_d \neq 0$ moduko aldebiko kontrasteak egiteko; aldez aurretiko jarraitu bat erabiltzen bada, ondorengoa ere jarraitua izango da, eta banaketa jarraitu orok puntu batean duen probabilitatea 0 da. Hori dela-eta, μ_d-ren sinesgarritasun-tartea erabili beharko da. 0 tartean badago, ezin da hipotesi nulua baztertu; 0 tartean ez badago, berriz, batezbestekoen diferentzia onartzen da $(1-\alpha) * 100$-eko probabilitatearekin.

4.2 Bariantza ezezaguna eta berdina eta aldez aurretiko laua duten bi banaketa normalen ausazko lagin independenteak

μ_1 eta μ_2-rako aldez aurretiko lau bat erabili denez gero, $(s'_1)^2 = \frac{\sigma^2}{n_1}$, $(s'_2)^2 = \frac{\sigma^2}{n_2}$, $m'_1 = \bar{y}_1$ eta $m'_2 = \bar{y}_2$ izango da.

4.2.1 Batezbestekoen diferentziarako sinesgarritasun-tarteak, bariantza ezezaguna eta berdina izanik

Bariantzaren balioa ezezaguna denez gero, datuetatik abiatuta kalkulatu beharko da.

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2}{n_1 + n_2 - 2}. \qquad (4.4)$$

Kasu horretan, sinesgarritasun-tarteak zabalagoak izango dira, bariantza zenbatestean ziurgabetasuna handiagoa baita. Balio kritikoa Studenten t banaketaren taulatik lortuko da ($n_1 + n_2 - 2$ askatasun-gradurekin). Hortaz, $\mu_1 - \mu_2$-ren sinesgarritasun-tarte bayestar hurbildua honako hau da:

$$\bar{y}_1 - \bar{y}_2 \pm t_{\frac{\alpha}{2}} * \hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \qquad (4.5)$$

4.2.2 Aldebakarreko hipotesi-test bayestarra

H_0: $\mu_d \leq 0$ versus H_1: $\mu_d > 0$ hipotesiak kontrastatu nahi badira (ausazko bi laginak $\hat{\sigma}_p^2$-ren arabera zenbatetsitako bariantza berdin ezezaguna duen normal batetik hartu direla eta μ_1 eta μ_2-ren aldez aurretiko banaketak lauak direla suposatuta), H_0-ren ondorengo probabilitatea kalkulatu beharko da (4.3) formularen bidez eta Studenten t banaketaren taula erabilita ($n_1 + n_2 - 2$ askatasun-gradurekin), normalaren taula erabili ordez.

4.2.3 Aldebiko hipotesi-test bayestarra

Baldintza beretan, (4.5) ekuazioan eman den eta $n_1 + n_2 - 2$ askatasun-gradu dituen $\mu_1 - \mu_2$ -ren sinesgarritasun-tartea erabiliko dugu H_0: $\mu_d = 0$ versus H_1: $\mu_d \neq 0$ kontrastatzeko. Berriz ere, α esangura-mailako hipotesi nulua baztertuko da 0a tartean ez badago.

4.3 Bariantza desberdin ezagunak dituzten bi banaketa normalen ausazko lagin independenteak

μ_1 batezbestekoa eta σ_1^2 bariantza ezaguna dituen banaketa normal batetik hartutako $y_{11}, \dots, y_{n_1 1}$ ausazko lagina eta μ_2 batezbestekoa eta σ_2^2 bariantza ezaguna dituen banaketa normal batetik hartutako $y_{12}, \dots, y_{n_2 2}$ ausazko lagina ditugu. Bi laginak elkarren independenteak dira.

Aldez aurretiko independenteak erabiliko ditugu μ_1 eta μ_2-rako, banaketa normalak edo lauak. Laginak eta banaketak elkarren independenteak direnez gero, a posteriori banaketak ere independenteak izango dira. Horretarako, (3.5) eta (3.6) ataletan emandako formulak erabiliko ditugu. Hortaz, $\mu_1|y_{11}, \dots, y_{n_1 1} \sim N(m'_1, (s'_1)^2)$-ren eta $\mu_2|y_{12}, \dots, y_{n_2 2} \sim N(m'_2, (s'_2)^2)$-ren ondorengo banaketak ditugu. Orduan, $\mu_d = \mu_1 - \mu_2$-ren ondorengo banaketa normala da.

$$\mu_d|y_{11}, \dots, y_{n_1 1} y_{12}, \dots, y_{n_2 2} \sim N(m'_d, (s'_d)^2)$$

Non $m'_d = m'_1 - m'_2$ eta $(s'_d)^2 = (s'_1)^2 + (s'_2)^2$.

4.3.1 Batezbestekoen diferentziarako sinesgarritasun-tarteak, bariantza desberdinak eta ezagunak izanik

Baldintza horietan, $(1 - \alpha) * 100$-eko sinesgarritasun-tarte bayestarra (4.1) eta (4.2) adierazpenen bidez lortzen da.

4.4 Bariantza desberdin ezezagunak dituzten bi banaketa normalen ausazko lagin independenteak

Bariantzak desberdinak eta ezezagunak direnean, honela zenbatetsi behar dira datuetatik abiatuta:

$$\hat{\sigma}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 \qquad y \qquad \hat{\sigma}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2$$

Zenbatesle horiek balio errealak baina ezezagunak balira bezala erabiliko dira eguneratze-formuletan. Horrek ziurgabetasuna gaineratzen duenez gero, Studenten t banaketaren taula erabili beharko da balio kritikoak kalkulatzeko. Kasu horretan, ez dago garbi erabili beharreko askatasun-graduen kopurua zein den, baina honako balio hau biribiltzea gomendatzen da (Satterthwaite):

$$\frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1 + 1} + \frac{(\hat{\sigma}_2^2/n_2)^2}{n_2 + 1}}. \qquad (4.6)$$

4.4.1 Batezbestekoen diferentziarako sinesgarritasun-tarteak, bariantza desberdinak eta ezezagunak izanik

Berriz ere (4.5) eta (4.6) formulak erabiliz, honela lortzen da $\mu_1 - \mu_2$-rako sinesgarritasun-tarte bayestar hurbildua:

$$m'_1 - m'_2 \pm t_{\frac{\alpha}{2}} * \sqrt{(s'_1)^2 + (s'_2)^2}\,.$$

Non (4.6) adierazpena hurbilen dagoen zenbaki osora biribiltzen den askatasun-graduak lortzeko. μ_1 eta μ_2-rako aldez aurretiko banaketa lau independenteak aukeratu direnean, aurreko adierazpena honela berridatz daiteke:

$$m'_1 - m'_2 \pm t_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}\ . \qquad (4.7)$$

4.4.2 Aldebakarreko hipotesi-test bayestarra

H_0: $\mu_d \leq 0$ versus H_1: $\mu_d > 0$ era bayestarrean α mailarekin kontrastatzeko, hipotesi nuluaren ondorengo probabilitatea kalkulatzen da. Horretarako, (5.3) ekuazioa erabiliko dugu, baina balio kritikoa normalaren taulatik lortu ordez, Studenten t banaketaren taulatik lortuko dugu, Satterthwaitek proposatutako askatasun-graduekin. Probabilitatea α baino txikiagoa bada, hipotesi nulua baztertu eta $\mu_1 > \mu_2$ ondorioztatzen da.

4.5 Proportzioen diferentziarako inferentzia bayestarra hurbilketa normal baten bidez

Demagun orain π_1 eta π_2 proportzioak konparatu nahi ditugula, sarrerako adibidean bezala. Demagun, halaber, $y|\pi_i$ banaketek $binomial(n_i, \pi_i)$ bati jarraitzen diotela eta elkarren independenteak direla ($i = 1,2$ izanik). Dakigunez, haien aldez aurretiko banaketak $beta(a_i, b_i)$ banaketak badira, ondorengoak n_i $beta(a_i^*, b_i^*)$ banaketak izango dira, non $a_i^* = a_i + y_i$ eta $b_i^* = b_i + y_i$ diren. Horrela, beraz, $\pi_d = \pi_1 - \pi_2$ diferentziaren ondorengo banaketa $N(m'_d, (s'_d)^2)$ normal baten bidez hurbil daiteke, non

$$m'_d = \frac{a_c^*}{a_c^* + b_c^*} - \frac{a_t^*}{a_t^* + b_t^*}, \quad (s'_d)^2 = \frac{a_c^* b_c^*}{(a_c^* + b_c^*)^2(a_c^* + b_c^* + 1)} + \frac{a_t^* b_t^*}{(a_t^* + b_t^*)^2(a_t^* + b_t^* + 1)}.$$

4.5.1 Proportzioen diferentziarako sinesgarritasun-tarteak

$\pi_d = \pi_1 - \pi_2$-rako $(1 - \alpha) * 100\%$-eko sinesgarritasun-tarte bayestar hurbildua honela lortzen da:

$$m'_d \pm z_{\frac{\alpha}{2}} * s'_d. \qquad (4.8)$$

4.5.2 Aldebakarreko hipotesi-test bayestarra

H_0: $\pi_d \leq 0$ versus H_1: $\pi_d > 0$ kontrastatu nahi baditugu, hipotesi nuluaren ondorengo banaketa kalkulatuko dugu honela:

$$P(\pi_d \leq 0|datos) = P\left(\frac{\pi_d - m'_d}{s'_d} \leq \frac{0 - m'_d}{s'_d}\right) = P\left(Z \leq \frac{0 - m'_d}{s'_d}\right). \qquad (4.9)$$

Probabilitatea α baino txikiagoa bada, hipotesi nulua baztertu eta $\pi_1 > \pi_2$ ondorioztatuko dugu.

4.5.3 Aldebiko hipotesi-test bayestarra

Aitzitik, H_0: $\pi_d = 0$ versus H_1: $\pi_d \neq 0$ kontrastatu nahi baditugu, (4.8) ekuazioak emandako sinesgarritasun-tartea aztertu beharko dugu. 0a tartean ez badago, hipotesi nulua baztertu eta lagin bakoitzaren proportzioak desberdinak direla onartuko da.

5. Erregresio linealerako inferentzia bayestarra

Zenbaitetan, interesgarria izaten da bi aldagairen (x eta y) arteko erlazioa modelatzea. Horretarako, iragartzen eta x-ek emandako informaziotik abiatzen lagunduko digun ekuazioa bilatzen da.

Demagun n datu-bikotez osatutako datu-base bat dugula, hots, $(x_i, y_i), i = 1, \dots, n$. Lehenik eta behin, komeni da *scatterplot* batean puntu guztiak marraztea, haien artean dagoen erlazioa intuitzen saiatzeko. Puntuak modurik sinpleenean erlazionatzen dituen ekuazioa $y = \alpha_0 + \beta x$ ekuazio lineala da. Ondoren, karratu txikienen metodoaz hitz egingo dugu. Izan ere, metodo hori da erabiliena α_0 eta zenbatesteko. β.

5.1 Karratu txikienak

Grafikoan, infinitu lerro marraz daitezke; batzuk ongi samar doituko dira puntuetara, eta beste batzuk, berriz, ez hain ongi. Hondarra puntu baten eta erregresio-lerroaren arteko distantzia bertikala da. Beraz, distantzia zenbat eta txikiagoa izan, orduan eta hobea izango da puntu zehatz horrekiko doikuntza.

Karratu txikienen metodoa hondarren karratuen batura minimizatuko duen zuzena aurkitzean eta oro har hobekien doitzen den zuzena lortzean datza.

5.1.1 Ekuazio normalak eta karratu txikienen zuzena

$\mathrm{y} = \alpha_0 + \beta\mathrm{x}$ adierazpeneko hondarren karratuen batura honako hau da:

$$SS_{res}(\alpha_0, \beta) = \sum_{i=1}^{n} [y_i - (\alpha_0 + \beta x_i)]^2. \qquad (5.1)$$

SS_{res} minimizatzen duten α_0 balioak eta β malda aurkitzeko, (6.1) ekuazioa α_0 eta β-rekiko deribatu, haiek 0rekin berdindu eta ekuazioen sistema ebazten da. Hau da,

$$\frac{\partial SS}{\partial \alpha_0} = \sum_{i=1}^{n} 2[y_i - (\alpha_0 + \beta x_i)]\,(-1) = 0 \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \alpha_0 - \sum_{i=1}^{n} \beta x_i = 0$$

$$\Downarrow$$

$$\bar{y} - n\alpha_0 - \beta\bar{x} = 0 \qquad (5.2)$$

Bestalde, β-rekiko deribatuz:

$$\frac{\partial SS}{\partial \beta} = \sum_{i=1}^{n} 2[y_i - (\alpha_0 + \beta x_i)]\,(-x_i) = 0 \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \alpha_0 x_i - \sum_{i=1}^{n} \beta {x_i}^2 = 0$$

$$\Downarrow$$

$$\overline{xy} - \alpha_0 \bar{x} - \beta \overline{x^2} = 0 \qquad (5.3)$$

(6.2) eta (6.3) ekuazioek ekuazio normalen sistema osatzen dute. (6.2) ekuaziotik abiatuta, α_0 β-ren arabera uzten da, eta (6.3) ekuazioan lortutako adierazpena aldatzen da.

$$\overline{xy} - (\bar{y} - \beta \bar{x})\bar{x} - \beta \overline{x^2} = 0 \Leftrightarrow$$

$$\Leftrightarrow \overline{xy} - \bar{y}\bar{x} + \beta \bar{x}\bar{x} - \beta \overline{x^2} = 0$$

$$\Leftrightarrow \beta\left(\bar{x}\bar{x} - \overline{x^2}\right) = \bar{y}\bar{x} - \overline{xy}$$

Hortaz, karratu txikienen maldak honako adierazpen hau dauka:

$$B = \frac{\bar{y}\bar{x} - \overline{xy}}{\left(\bar{x}\bar{x} - \overline{x^2}\right)} \qquad (5.4)$$

(6.2) ekuaziora itzuliz eta maldaren balioa aldatuz, honako hau dugu:

$$A_0 = \bar{y} - B\bar{x} \qquad (5.5)$$

Hortaz, karratu txikienen bidez lortutako zuzena $y = A_0 + Bx$ da.

5.1.2 Karratu txikienen zuzenaren alternatiba

Maldak eta y interzeptuaren gaineko beste edozein puntuk ere zuzena finkatzeko aukera ematen digute. Demagun $A_{\bar{x}}$ karratu txikienen zuzenak -en zuzen bertikala mozten duen puntua dela. $\bar{x}$.

$$A_{\bar{x}} = A_0 + B\bar{x} = \bar{y}$$

Hau da, karratu txikienen zuzena $(\bar{x}, \bar{y})$ puntutik igarotzen da, eta oso baliagarria izango zaigun honako ekuazio hau lor daiteke:

$$y = A_{\bar{x}} + B(x - \bar{x}) = \bar{y} + B(x - \bar{x}) \qquad (5.6)$$

5.1.3 Karratu txikienen zuzenaren inguruko bariantzaren estimazioa

Hona hemen karratu txikienen zuzenaren inguruko bariantzaren estimazioa:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}[y_i - (A_{\bar{x}} + B(x_i - \bar{x}))]^2}{n-2},$$

Beraz, hondarren karratuen batura da, $n-2$-rekin zatitua, $A_{\bar{x}}$ eta B-ren estimazioak erabiltzearen ondorioz.

5.1.4 Adibidea

Enpresa batean, produktu bat salgai jartzen denean zer hezetasun-maila duen jakin nahi dute. Dena dela, produktua burututa dagoenean ez da ekonomikoki bideragarria hezetasuna neurtzea. Horregatik, hezetasuna produktua amaitu aurreko fasean neurtzea eta balio horretatik abiatuta amaierako hezetasun-maila aurresatea pentsatu dute. Bildu diren 25 laginetan, x aldagaiak aldez aurretiko hezetasun-maila neurtzen du, eta y aldagaiak, berriz, amaierako hezetasun-maila. Honako zenbakizko emaitza hauek lortu dira: $\bar{x} = 14.3888$, $\bar{y} = 14.2208$, $\overline{x^2} = 207.0703$, $\overline{y^2} = 202.3186$ eta $\overline{xy} = 204.6628$.

Karratu txikienen erregresio-zuzena lortzeko, maldaren balioa behar dugu, eta honela lortzen da:

$$B = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{204.6628 - 14.3888 * 14.2208}{207.0703 - (14.3888)^2} = 1.29963.$$

Hortaz, karratu txikienen zuzena honako hau da: $y = 14.2208 + 1.29963 * (x - 14.2208)$.

$\hat{y}_i = \bar{y} + B(x_i - \bar{x})$ balio doituak, $y - \hat{y}$ hondarrak eta haien karratuak ($i = 1, \ldots, 25$ izanik) lortzen dira. Hortaz, karratu txikienen bidez zenbatetsitako bariantza eta desbiderapena honako hauek dira:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{0.801882}{23} = 0.0348644 \implies \hat{\sigma} = \sqrt{0.034864} = 0.18672$$

5.2 Erregresio lineal sinplearen hipotesia

Karratu txikienen metodoa teknika ez-parametrikotzat hartzen da, ez baitu datuen banaketa erabiltzen. Maldaren eta interzeptuaren balioan inferitzeko, hipotesi batzuk hartu behar dira.

Parametrizazio alternatiboa erabiliz gero, honako hau lortuko dugu:

$$y_i = \alpha_{\bar{x}} + \beta(x_i - \bar{x}) + e_i,$$

non $\alpha_{\bar{x}}$ y-ren batez besteko balioa den ($x = \bar{x}$ izanik), eta β malda den. e_i errore bakoitza normal banatuta dago 0 batezbestekoarekin eta σ^2 bariantza ezagunarekin, eta erroreek bariantza konstantea dute. Gainera, erroreak elkarren independenteak dira. Hortaz, $y_i|x_i$-k banaketa normala dauka, $\alpha_{\bar{x}} + \beta(x_i - \bar{x})$ batezbestekoarekin eta σ^2 bariantzarekin, eta $y_i|x_i$ guztiak elkarren independenteak dira ($i = 1, .. n$ izanik).

5.3 Bayesen teorema erregresio-eredurako

Bayesen teorema honela laburbil daiteke:

$$posterior \propto previa * verosimilitud,$$

Horregatik, egiantza finkatu eta eredurako aldez aurretikoa erabaki behar da.

5.3.1 β eta -ren baterako egiantza $\alpha_{\bar{x}}$

i-garren behaketaren baterako egiantza haren dentsitate-funtzioa da, $\alpha_{\bar{x}}$ eta β parametroen funtzio gisa ikusita, non (x_i, y_i) balio finkoak diren. Hortaz, parametrorik ez duen zatia alde batera utzita, i behaketaren egiantza funtzio honen bidez lortzen da: L:

$$L_i(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}[y_i-(\alpha_{\bar{x}}+\beta(x_i-\bar{x}))]^2}$$

Behaketak elkarren independenteak direnez gero, lagin osoaren egiantza behaketa guztien egiantzen biderkadura da.

$$L_{muestra}(\alpha_{\bar{x}}, \beta) \propto \prod e^{-\frac{1}{2\sigma^2}[y_i-(\alpha_{\bar{x}}+\beta(x_i-\bar{x}))]^2}.$$

Esponentzialaren propietateak aplikatuz gero,

$$L_{muestra}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}[y_i-(\alpha_{\bar{x}}+\beta(x_i-\bar{x}))]^2]}.$$

Kako zuzenen arteko adierazpena honela berridatz daiteke:

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

$$\Downarrow$$

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

$SS_y = \sum_{i=1}^{n}(y_i-\bar{y})^2$, $SS_{xy} = \sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$ eta $SS_x = \sum_{i=1}^{n}(x_i-\bar{x})^2$ hartuta, aurreko adierazpena honela sinplifikatzen da:

$$SS_y - 2\beta SS_{xy} + \beta^2 SS_x + n(\alpha_{\bar{x}}-\bar{y})^2.$$

Hortaz, baterako egiantza honela idatz daiteke:

$$L_{muestra}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}[SS_y-2\beta SS_{xy}+\beta^2 SS_x+n(\alpha_{\bar{x}}-\bar{y})^2]},$$
$$L_{muestra}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}[SS_y-2\beta SS_{xy}+\beta^2 SS_x]} * e^{-\frac{1}{2\sigma^2}[n(\alpha_{\bar{x}}-\bar{y})^2]}.$$

Lehen gaian SS_x faktore komun gisa ateratzen, biderkadura nabarmenaren formulak aplikatzen eta ezein parametroren mende ez dauden gaiak alde batera uzten baditugu,

$$L_{muestra}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2/SS_x}\left[\beta-\frac{SS_{xy}}{SS_x}\right]} * e^{-\frac{1}{2\sigma^2/n}[(\alpha_{\bar{x}}-\bar{y})^2]}.$$

Kontuan hartu behar da $\frac{SS_{xy}}{SS_x} = B$ dela (karratu txikienen bidez lortutako malda), eta $\bar{y} = A_{\bar{x}}$ (karratu txikienen bidez zenbatetsitako interzeptua $x = \bar{x}$ lerro bertikalean). Hortaz, bi egiantzen biderkaduraren bidez lortzen da baterako egiantza:

$$L_{muestra}(\alpha_{\bar{x}}, \beta) \propto L_{muestra}(\alpha_{\bar{x}}) * L_{muestra}(\beta),$$

Non

$$L_{muestra}(\alpha_{\bar{x}}) \propto e^{-\frac{1}{2\sigma^2/SS_x}[\beta - B]} \text{ eta } L_{muestra}(\beta) \propto e^{-\frac{1}{2\sigma^2/n}[(\alpha_{\bar{x}} - A_{\bar{x}})^2]}.$$

Baterako egiantza bi egiantz independenteren biderkadurarekin faktorizatu da, eta, beraz, egiantz indibidualak independenteak dira. Ikus daitekeenez, β maldaren egiantza B batezbestekoa eta $2\sigma^2/SS_x$ bariantza dituen normal bat da, eta $\alpha_{\bar{x}}$-ren egiantza normala da, $A_{\bar{x}}$ batezbestekoarekin eta σ^2/n bariantzarekin.

5.3.2 β eta -ren aldez aurretiko baterako banaketa $\alpha_{\bar{x}}$

Baterako egiantza baterako aldez aurretikoarekin biderkatzen bada, biderkadura ondorengo baterakoaren proportzionala izango da. Aldez aurretiko baterakoa bi aldez aurretiko independente eta indibidualen arteko biderkadura da, eta bi aldez aurretiko horiek normalak edo lauak izan daitezke.

$$g(\alpha_{\bar{x}}, \beta) = g(\alpha_{\bar{x}}) * g(\beta),$$

5.3.2.1 Aldez aurretiko normala

$\alpha_{\bar{x}}$-rako aldez aurretiko batezbestekoa honela lortzen da: $m_{\alpha_{\bar{x}}}$. Ondoren, y,-k har ditzakeen balio maximoa eta minimoa hartuko dira, eta $s_{\alpha_{\bar{x}}}$ aldez aurretiko desbiderapena bi balio horien diferentzia 6rekin zatituta lortuko da.

Malda nulua izan daitekeela kontuan hartuta, β-ren aldez aurretiko batezbestekoa $m_\beta = 0$. dela finkatzen da. Arestian azaldutako prozedura erabiliz, s_β kalkulatuko litzateke, eta gure aldez aurretiko ezagutza honela adieraziko litzateke: $N(m_\beta, s_\beta^2)$.

5.3.3 β eta -ren ondorengo baterako banaketa $\alpha_{\bar{x}}$

Ondorengo baterako banaketa aldez aurretiko banaketaren eta baterako egiantzaren biderkaduraren proportzionala da.

$$g(\alpha_{\bar{x}}, \beta|\text{data}) \propto g(\alpha_{\bar{x}}, \beta) * L_{muestra}(\alpha_{\bar{x}}, \beta),$$

Adierazpen horretan, $(x_1, y_1), \ldots, (x_n, y_n)$ bikote ordenatuen multzoa da data. Ondorengo marjinalak erabiliz gero, honela berridatz dezakegu ekuazioa:

$$g(\alpha_{\bar{x}}, \beta|\text{data}) \propto g(\alpha_{\bar{x}}|\text{data}) * g(\beta|\text{data}).$$

Ondorengo marjinal bakoitza banaketa normaletarako eguneratze-arauen bidez lor daiteke. Are gehiago, β-rako $N(m_\beta, s_\beta^2)$ aldez aurretiko banaketa batetik abiatuta, $N(m'_\beta, (s'_\beta)^2)$ ondorengo banaketa lortuko dugu, non

$$\frac{1}{(s'_\beta)^2} = \frac{1}{s_\beta^2} + \frac{SS_x}{\sigma^2} \qquad (5.7)$$

$$m'_{\beta} = \frac{\frac{1}{s_{\beta}^2}}{\frac{1}{(s'_{\beta})^2}} * m_{\beta} + \frac{\frac{SS_x}{\sigma^2}}{\frac{1}{(s'_{\beta})^2}} * B \qquad (5.8)$$

Ondorengo zehaztasuna aldez aurretiko zehaztasunaren eta egiantzaren zehaztasunaren arteko baturaren baliokidea da. Ondorengo batezbestekoa aldez aurretiko batezbestekoaren eta egiantzaren batezbestekoaren haztapen bat da, ondorengo zehaztasunaren proportzionalak diren zehaztasun-pisuekin. Hortaz, ondorengo banaketa ere normala da. $\alpha_{\bar{x}}$-rako $N(m_{\alpha_{\bar{x}}}, s^2_{\alpha_{\bar{x}}})$ aldez aurretiko banaketa batetik abiatuta, $N(m'_{\alpha_{\bar{x}}}, (s'_{\alpha_{\bar{x}}})^2)$ ondorengo banaketa lortuko dugu, non

$$\frac{1}{(s'_{\alpha_{\bar{x}}})^2} = \frac{1}{s^2_{\alpha_{\bar{x}}}} + \frac{SS_x}{\sigma^2} \qquad (5.9)$$

$$m'_{\alpha_{\bar{x}}} = \frac{\frac{1}{s^2_{\alpha_{\bar{x}}}}}{\frac{1}{(s'_{\alpha_{\bar{x}}})^2}} * m_{\alpha_{\bar{x}}} + \frac{\frac{n}{\sigma^2}}{\frac{1}{(s'_{\alpha_{\bar{x}}})^2}} * A_{\bar{x}} \qquad (5.10)$$

5.3.4 Adibidea (jarraipena)

Aurreko enpresako estatistikariak erabaki du $N(1, (0.3)^2)$ bat erabiltzea β-ren aldez aurretiko gisa, eta $N(15, 1^2)$ bat $\alpha_{\bar{x}}$-ren aldez aurretiko gisa. Bariantza erreala ezezaguna denez gero, karratu txikienen bidez lortutako estimazioa erabiliko du: $\hat{\sigma}^2 = 0.0348644$.

Kontuan hartu behar da $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n(\overline{x^2} - \bar{x}^2) = 25 * (207.0703 - 14.3888^2) = 0.81886$ dela. Hortaz, β-ren ondorengo zehaztasuna eta bariantza eta ondorengo batezbestekoa honako hauek dira:

$$\frac{1}{(s'_{\beta})^2} = \frac{1}{0.3^2} + \frac{0.811886}{0.0348644} = 34.5981 \implies s'_{\beta} = 0.17001.$$

$$m'_{\beta} = \frac{\frac{1}{0.3^2}}{34.5981} * 1 + \frac{\frac{0.81886}{0.0348644}}{34.5981} * 1.29963 = 1.2034.$$

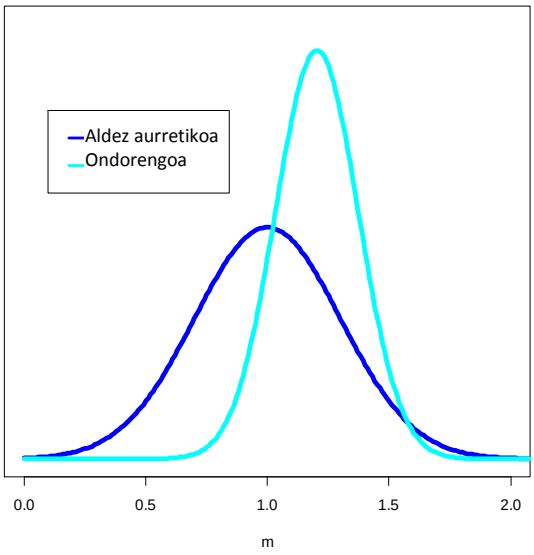
Era berean, $\alpha_{\bar{x}}$-ren ondorengo balioak lortuko ditugu:

$$\frac{1}{(s'_{\alpha_{\bar{x}}})^2} = \frac{1}{1^2} + \frac{25}{0.0348644} = 718.064 \implies s'_{\alpha_{\bar{x}}} = 0.037318.$$

$$m'_{\alpha_{\bar{x}}} = \frac{\frac{1}{1^2}}{718.064} * 15 + \frac{\frac{25}{0.0348644}}{718.064} * 14.2208 = 14.2219.$$

Ondorengo grafikoan, maldaren aldez aurretiko eta ondorengo banaketak ditugu.

Maldaren banaketak



5.3.4 Maldarako sinesgarritasun-tartea

β-ren ondorengo banaketak datuak behatu ondoren dugun ezagutza adierazten du. β,-rako sinesgarritasun-tarte bayestar bat lortu nahi bada, honela adieraziko da:

$$m'_\beta \pm z_{\frac{\alpha}{2}}\sqrt{\left(s'_\beta\right)^2} \qquad (5.11)$$

Hala ere, errealistagoa da pentsatzea ezezaguna izango dela σ^2-ren balioa. Hondarretatik abiatuta, honako estimazio hau egin daiteke:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - (A_{\bar{x}} + B(x_i - \bar{x})))^2}{n-2}.$$

Ziurgabetasun-maila handiagoa denez gero, tarteak zabalagoa izan behar du. Horretarako, estatistiko normalaren balio kritikoaren ordez, Studenten t banaketarena erabiliko dugu, $n-2$ askatasun-gradurekin. Hortaz, tarteak honako adierazpen hau izango du:

$$m'_\beta \pm t_{\frac{\alpha}{2}}\sqrt{\left(s'_\beta\right)^2} \qquad (5.12)$$

5.3.5 Maldarako aldebakarreko kontrastea

Askotan, x-eko unitate bateko gehikuntzari lotutako y -ko gehikuntza β_0 balio jakin bat baino handiagoa den jakin nahi izaten da. Horretarako, honako hipotesi hauek planteatzen dira:

$$H_0: \beta \leq \beta_0 \text{ vs. } H_1: \beta > \beta_0$$

Hipotesiak era bayestarrean eta α mailan testatzeko, hipotesi nuluaren ondorengo probabilitatea kalkulatzen da:

$$P(\beta \leq \beta_0|data) = \int_{-\infty}^{\beta_0} g(\beta|data)d\beta = P\left(Z \leq \frac{\beta_0 - m'_\beta}{s'_\beta}\right). \quad (5.13)$$

Probabilitate hori α baino txikiagoa bada, hipotesi nulua baztertu egiten da. Bariantza zenbatetsi behar izatekotan, Z normalaren ordez Studenten t banaketa erabili beharko da, $n-2$ askatasun-gradurekin.

5.3.6 Maldarako aldebiko kontrastea

Baldin eta $\beta = 0$ bada, y-ren batezbestekoa ez dago x-en mende. Beraz, interesgarria da H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$ hipotesiak era bayestarrean eta α esangura-mailarekin kontrastatzea aurresateak egiteko erregresio-eredua erabili aurretik. Horretarako, 0 balioa sinesgarritasun-tartean ote dagoen egiaztatu behar da. Tarte horretan ez badago, hipotesi nulua baztertu egiten da, eta egokia da erregresioa eredu aurresale gisa erabiltzea.

5.4 Etorkizuneko behaketetarako banaketa aurresaleak

Erregresio linealaren helburu nagusietako bat y_{n+1} balioa aurresatea da, x_{n+1} aldagai azaltzailearen balioa emanik eta mendeko aldagaiaren eta aldagai azaltzailearen artean benetako erlazio lineal bat dagoela egiaztatu ondoren. Hortaz, y_{n+1}-en aurresaterik onena, x_{n+1} emanik, honako hau izango da:

$$\tilde{y}_{n+1} = \hat{\alpha}_{\bar{x}} + \hat{\beta}(x_{n+1} - \bar{x}),$$

Adierazpen horretan, $\hat{\beta}$ zenbatetsitako malda da, eta $\hat{\alpha}_{\bar{x}}$, berriz, $x = \bar{x}$ zuzenaren interzeptuaren estimazioa.

Aurresatearen egokitasunari dagokionez, bi ziurgabetasun-puntu daude. Lehenik eta behin, balio erreal baina ezezagunaren ordez bi balioren estimazioa erabili da aurresatea egiteko. Parametroak ausazko aldagaiak direla eta haien ondorengo banaketa aurreko sekzioan lortu dela suposatzen da. Gainera, y_{n+1} behaketa berriak bere errore propioa dauka, eta aurreko erroreekiko independentea izango da. y_{n+1}-en banaketa aurresalea, x_{n+1} eta datuak emanik, Bayesen teorema aplikatuta lortzen da, eta honela adierazten da: $f(y_{n+1}|x_{n+1}, data)$.

5.4.1 Banaketa aurresalea

Banaketa aurresalea honako integral honetatik abiatuta lortzen da:

$$f(y_{n+1}|x_{n+1}, data) = \iint f(y_{n+1}, \alpha_{\bar{x}}, \beta|x_{n+1}, data)d\alpha_{\bar{x}}d\beta.$$

Lehenik eta behin, parametroen eta hurrengo behaketaren ondorengo baterakoa finkatu behar da, x_{n+1} balioa eta datuak emanik:

$$(y_{n+1}, \alpha_{\bar{x}}, \beta|x_{n+1}, data) = f(y_{n+1}|\alpha_{\bar{x}}, \beta, x_{n+1}, data) * g(\alpha_{\bar{x}}, \beta|x_{n+1}, data).$$

Hurrengo y_{n+1} behaketa, $\alpha_{\bar{x}}$ eta β parametroak eta x_{n+1}-en balioa emanik, erregresio-ereduarekin lortutako ausazko aldagai berri bat da. $\alpha_{\bar{x}}$ eta β emanik, behaketa guztiak elkarren independenteak dira. Hau da, parametroak emanik, behaketa berria ez dago aldez aurretiko aurresateez osatutako datuen mende. Behaketa berriaren eta parametroen baterako banaketa honela sinplifikatzen da:

$$f(y_{n+1}, \alpha_{\bar{x}}, \beta | x_{n+1}, data) = f(y_{n+1} | \alpha_{\bar{x}}, \beta, x_{n+1}) * g(\alpha_{\bar{x}}, \beta | data)$$

Hartutako hipotesien arabera, hurrengo behaketa normal bat bezala banatzen da, $\mu_{n+1} = \alpha_{\bar{x}} + \beta(x_{n+1} - \bar{x})$ batezbestekoarekin eta σ^2 bariantza ezagunarekin. Parametroen ondorengo banaketak, aurreko sekzioan ikusitako eguneratze-arauetatik abiatuta lortutako aldez aurretiko datuak emanik, $N(m'_{\alpha_{\bar{x}}}, (s'_{\alpha_{\bar{x}}})^2)$ eta $N(m'_{\beta}, (s'_{\beta})^2)$ dira, independenteak. Behaketa berria $\mu_{n+1} = \alpha_{\bar{x}} + \beta(x_{n+1} - \bar{x})$, funtzio linealaren bitartez soilik dago parametroen mende, eta, beraz, arazoa sinplifikatu egin daiteke. $\alpha_{\bar{x}}$ eta β independenteak direnez gero, μ_{n+1}-en ondorengo banaketa normala izango da, $m'_{\mu} = m'_{\alpha_{\bar{x}}} + (x_{n+1} - \bar{x}) * m'_{\beta}$ batezbestekoarekin eta $(s'_{\mu})^2 = (s'_{\alpha_{\bar{x}}})^2 + (x_{n+1} - \bar{x})^2 * (s'_{\beta})^2$ bariantzarekin.

Banaketa aurresalea lortzeko, μ_{n+1}-en marjinalizazioa erabiliko dugu y_{n+1} eta μ_{n+1}-en ondorengo baterakoaren bidez.

$$f(y_{n+1} | x_{n+1}, data) = \int f(y_{n+1}, \mu_{n+1} | x_{n+1}, data) d\mu_{n+1}$$

$$= \int f(y_{n+1} | \mu_{n+1}, x_{n+1}, data) * g(\mu_{n+1} | x_{n+1}, data) d\mu_{n+1}$$

$$= \int f(y_{n+1} | \mu_{n+1}) * g(\mu_{n+1} | x_{n+1}, data) d\mu_{n+1}$$

$$\propto \int e^{-\frac{1}{2\sigma^2}(y_{n+1} - \mu_{n+1})^2} * e^{-\frac{1}{2(s'_{\mu})^2}(\mu_{n+1} - m'_{\mu})^2} d\mu_{n+1}$$

$$\propto \int e^{-\frac{1}{\frac{2\sigma^2(s'_{\mu})^2}{\sigma^2 + (s'_{\mu})^2}}\left(\mu_{n+1} - \frac{y_{n+1}(s'_{\mu})^2 + m'_{\mu}\sigma^2}{\sigma^2 + (s'_{\mu})^2}\right)^2} * e^{-\frac{1}{2\left((s'_{\mu})^2 + \sigma^2\right)}(y_{n+1} - m'_{\mu})^2} d\mu_{n+1}$$

Bigarren faktorea ez dago μ_{n+1}-en mende, eta, lehenengoaren integrala sinplifikatuz gero,

$$f(y_{n+1} | x_{n+1}, data) \propto e^{-\frac{1}{2\left((s'_{\mu})^2 + \sigma^2\right)}(y_{n+1} - m'_{\mu})^2}.$$

Hortaz, $N(m'_y, (s'_y)^2)$ bat dugu, non $m'_y = m'_{\mu}$ eta $(s'_y)^2 = \left((s'_{\mu})^2 + \sigma^2\right)$ den.

5.4.2 Aurresateko sinesgarritasun-tarteak

Interesgarria izaten da y_{n+1} balioa edukitzeko $1 - \alpha$-ko ondorengo probabilitatea duen tarte bat bilatzea, x_{n+1} balioan behatuta. Aurresatearen m'_y batezbestekotik eta $(s'_y)^2$ bariantzatik abiatuta, sinesgarritasun-tartea honela lortzen da:

$$m'_y \pm z_{\frac{\alpha}{2}} * s'_y = m'_{\mu} \pm z_{\frac{\alpha}{2}} * \sqrt{(s'_{\mu})^2 + \sigma^2}$$

$$= m'_{\alpha_{\bar{x}}} + (x_{n+1} - \bar{x}) * m'_{\beta} \pm z_{\frac{\alpha}{2}} * \sqrt{(s'_{\alpha_{\bar{x}}})^2 + (x_{n+1} - \bar{x})^2 * (s'_{\beta})^2 + \sigma^2},$$

behaketaren σ^2 bariantza ezaguna denean. Bariantza ezezaguna delako zenbatetsi behar bada, berriz,

$$m'_y \pm t_{\frac{\alpha}{2}} * s'_y = t'_\mu \pm z_{\frac{\alpha}{2}} * \sqrt{\left(s'_\mu\right)^2 + \sigma^2}$$

$$= m'_{\alpha_{\bar{x}}} + (x_{n+1} - \bar{x}) * m'_\beta \pm t_{\frac{\alpha}{2}} * \sqrt{\left(s'_{\alpha_{\bar{x}}}\right)^2 + (x_{n+1} - \bar{x})^2 * \left(s'_\beta\right)^2 + \sigma^2},$$

non balio kritikoa Studenten t banaketaren bidez lortzen den, $n-2$ askatasun-gradurekin. Kasu horretan, tarte bayestarrak eta frekuentistak berdinak dira.

5.4.3 Adibidea (jarraipena)

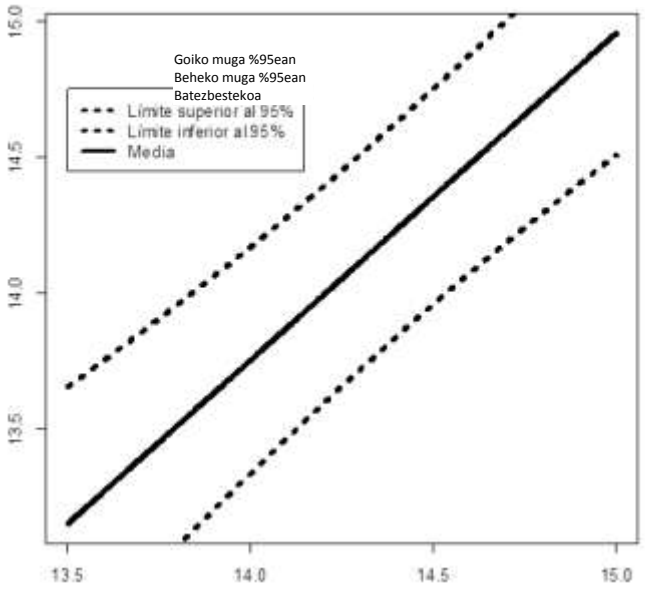
Bitarteko mailaren (x) araberako amaierako hezetasun-mailaren (y) banaketa aurresalea kalkulatzen da. Banaketa aurresalearen batezbestekoa eta bariantza honako hauek dira:

$m'_y = 14.2219 + 1.2034(x - 14.3888)$,

$$\left(s'_y\right)^2 = 0.0348644 + 0.037318^2 + 0.17001^2 * (x - 14.3888)^2.$$

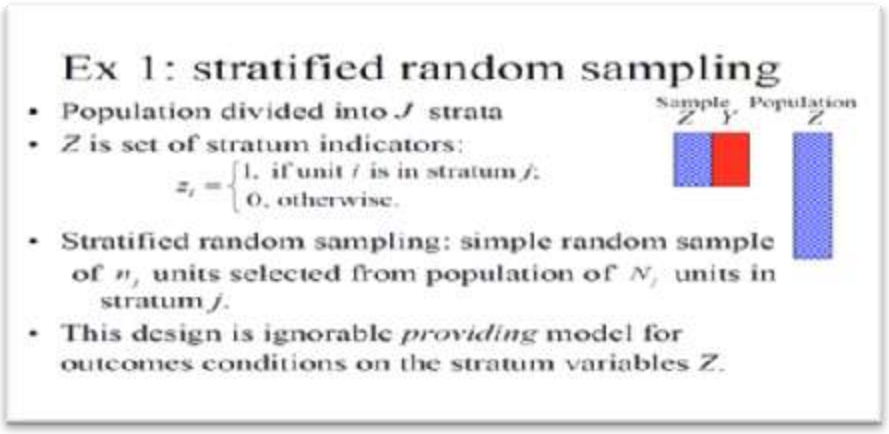
Hortaz, 95%eko sinesgarritasun-tartea $(m'_y - t_{0.025} * s'_y, m'_y + t_{0.025} * s'_y)$ da. Ondorengo grafikoan, batezbestekoa eta % 95eko tarte bayestarra daude irudikatuta.

Batezbesteko aurresalea, beheko eta goiko mugekin



6. Simulazio bidezko adibideak

Kapitulu honetan, ausazko lagin estratifikatu batean aplikatuko dugu landutako teoria, eta, ondoren, lortutako emaitzak aztertuko ditugu. Lehenik eta behin, batezbesteko normal baten estimazio estratifikatuarekin lan egingo dugu, eta, ondoren, proportzio baten estimazio estratifikatuarekin. Bi adibideetan, egiantzeko hainbat egoera adieraziko dituzten datu simulatuekin lan egingo dugu.



6.1 Adibidea batezbesteko normal batekin

Demagun N tamainako populazio batetik hartutako n tamainako ausazko lagin bat dugula eta ezaugarri jakin baten balioa neurtu nahi dugula. Laginaren behaketak bi estratutan banatuta daude, eta estratu bakoitza $\mathbf{z} = (\mathrm{z}_1, \mathrm{z}_2)$ bektore dikotomikoaren bidez adierazten da, non $\mathrm{z}_j = 1$ek adierazten duen behaketa j estratukoa dela, j $\{1,2\}$. izanik. Kasu horretan, alde batetik $\mathrm{n}_1 = 250$ eta $\mathrm{n}_2 = 150$ $(\mathrm{n} = 400)$ dugu, eta, bestetik, $\mathrm{N}_1 = 5000$ eta $\mathrm{N}_2 = 3000$ $(\mathrm{N} = 8000)$. Helburua estratu bakoitzeko Y aldagaiaren a posteriori batezbestekoa eta aldagaiaren guztizko batezbestekoa aurkitzea izango da.

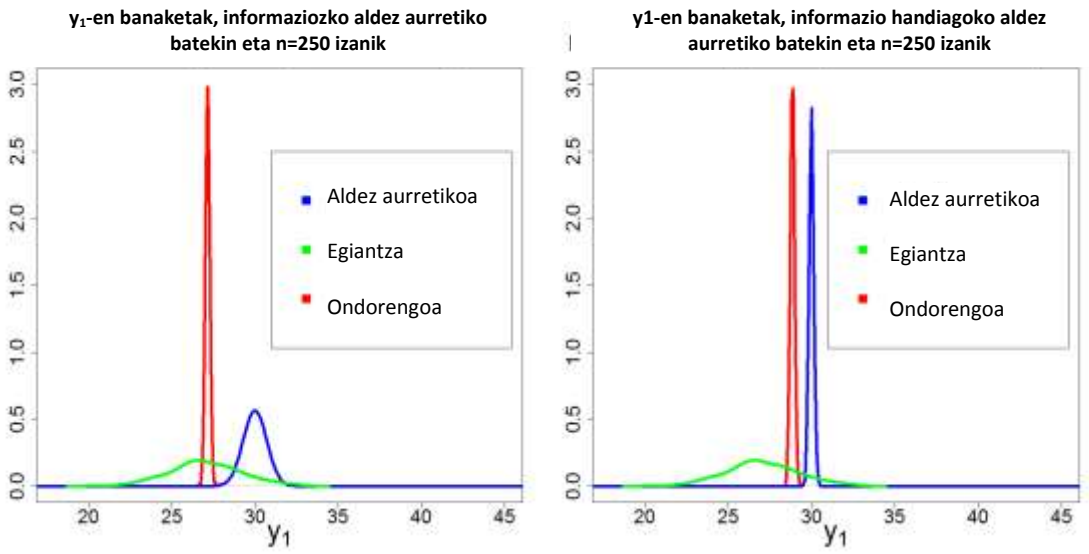
Demagun, halaber, aurreko azterlanei erreparatuz badakigula Y aldagaia normal banatuta dagoela estratu bakoitzean, eta estratu bakoitzaren batezbestekoak lehen estratuan 30eko batezbestekoa eta 0.5eko bariantza, eta bigarrenean 300eko batezbestekoa eta 10eko bariantza dituen normal bati jarraitzen diola. Y-ren bariantzari buruzko informaziorik ez dugu.

Hortaz, $\boldsymbol{z} = (1,0)$ betetzen duten behaketei (lehen estratukoei) dagokienez, pentsa dezakegu $\mathrm{Y}_1 \sim \mathrm{N}(\mathrm{mu}, \mathrm{tau})$ dela (non tau bariantzaren alderantzizkoa den) eta, era berean, $\mathrm{mu} \sim \mathrm{N}(30, \frac{1}{0.5})$ dela. Aldez aurretiko bariantzari buruzko a priori informaziorik ez dugunez gero, informaziozkoa ez den aldez aurretiko bat esleituko diogu tau-ri: adibidez, $\mathrm{tau} \sim \mathrm{Ga}(0.001, 0.001)$. Aldez aurretikoari indar handiagoa eman nahi badiogu, bariantza txikiagotu dezakegu. Adibidez, pentsa liteke $\mathrm{mu} \sim \mathrm{N}(30, \frac{1}{0.02})$ dela.

Aldez aurretiko banaketa finkatu ondoren, behatutako datuek ematen diguten informazioa ikusi beharko litzateke. Adibide honetarako, lehen estratuko 250 behaketak simulatu ditugu – $\mathrm{N}(27,\frac{1}{5})$ banaketa bati jarraitzen diotela suposatuta–, eta 27.05eko lagin-batezbestekoa lortu dugu, 4.63ko bariantzarekin.

Hirugarren eta azken urratsa a posteriori banaketa lortzea izango da. Eskura ditugun datuekin eguneratze-araurik erabili ezin denez gero, ondorengo banaketa simulatuko dugu MCMC metodoen bidez. Horretarako, WinBUGS programa erabiliko dugu, R-tik deituta. Erabilitako kodea I. Eranskinean agertzen da.

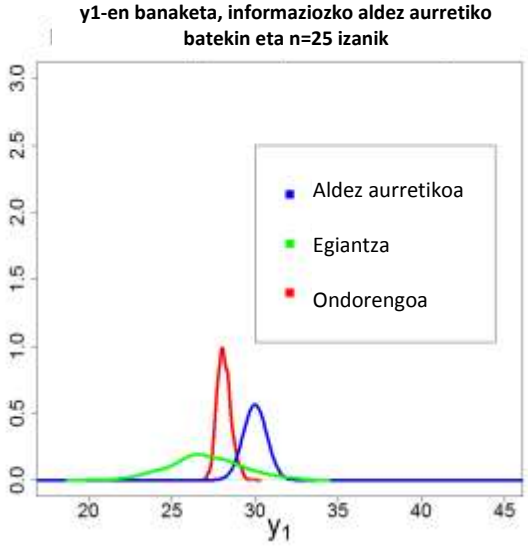
Ondorengo grafikoetan, proposatutako bi aldez aurretikoetarako lortutako banaketak ditugu:

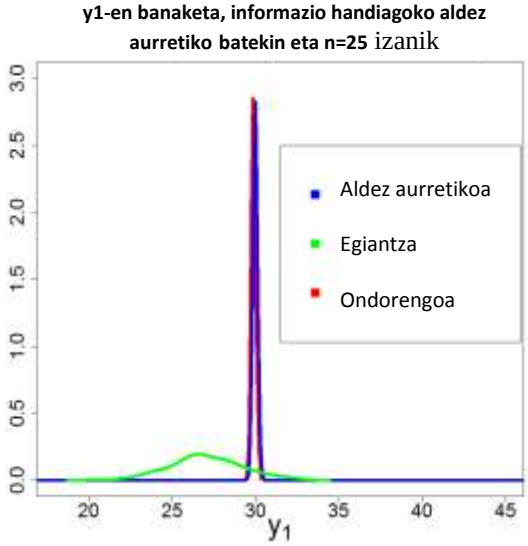


Estimazio puntual gisa, a posteriori batezbestekoa erabiliko dugu: $y_{1.1} = 27.16$ eta $y_{1.2} = 28.86$. Lehen kasuan, a posteriori batezbestekoa laginaren batezbestekotik oso hurbil dago. Horrek esan nahi du aldez aurretikoak ia ez duela garrantzirik eta 250 behaketaren informazioarekin osatutako egiantzak baino askoz pisu txikiagoa duela. Bigarren kasuan, a posteriori batezbestekoa laginaren batezbestekoaren eta a priori batezbestekoaren erdialdean dago. Informazio handiagoko aldez aurretiko bat ezarri denez gero, pisu handiagoa ematen zaio, eta egiantzaren bidez lortutako informazioa bezain garrantzitsua da.

Demagun orain laginaren tamaina 250 behaketatik 25era murrizten dugula eta datuak $\mathrm{N}(27,\frac{1}{5})$ banaketa bera erabiliz simulatzen ditugula; laginaren batezbestekoa 27.38 izango da, eta bariantza, berriz, 4.51.

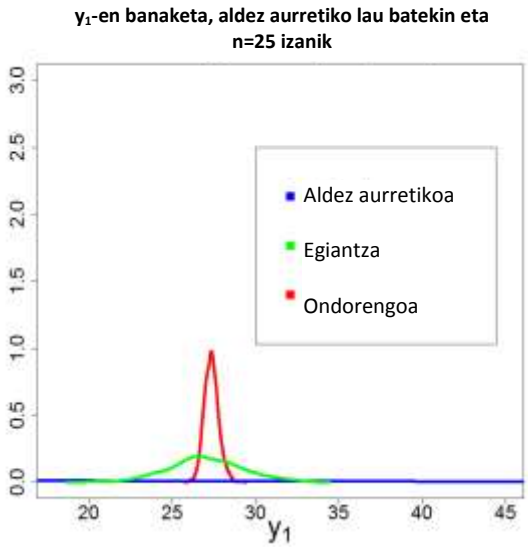
Arestian mu, $\mathrm{mu}\sim\mathrm{N}(30,\frac{1}{0.5})$ eta $\mathrm{mu}\sim\mathrm{N}(30,\frac{1}{0.02})$-rako proposatutako bi aldez aurretikoak erabiltzen baditugu, honako grafiko hauek lortuko ditugu:





Kasu horretan, a posteriori banaketak zein a posteriori batezbestekoaren estimazio puntualak ($y_{1.1} = 28.15$ eta $y_{1.2} = 29.88$,) laginaren tamaina handiagoa zenean lortutakoenen desberdin samarrak dira. Aldez aurretiko lehenarekin, aldez aurretiko batezbestekoaren eta laginaren batezbestekoaren artean dagoen estimazio bat lortzen da. Horrek esan nahi du aukeratutako aldez aurretikoak eta egiantzak antzeko garrantzia dutela eta, laginaren tamaina murriztu izanaren ondorioz, aldez aurretiko ezagutza erabiltzen dela emaitza eguneratuak lortzeko. Bigarren kasuan, balio berean zentratuta dagoen, baina bariantza txikiagoa duen aldez aurretiko bat erabiltzen da. A posteriori batezbestekoaren estimazioa a priori batezbestekotik oso hurbil dago; izan ere, lagin-datu gutxi (n=25) daudenez eta aldez aurretikoa oso sendoa denez gero, egiantza ez da kontuan hartzen, eta ia aldez aurretiko banaketaren informazioa baino ez da erabiltzen estimazioa egiteko.

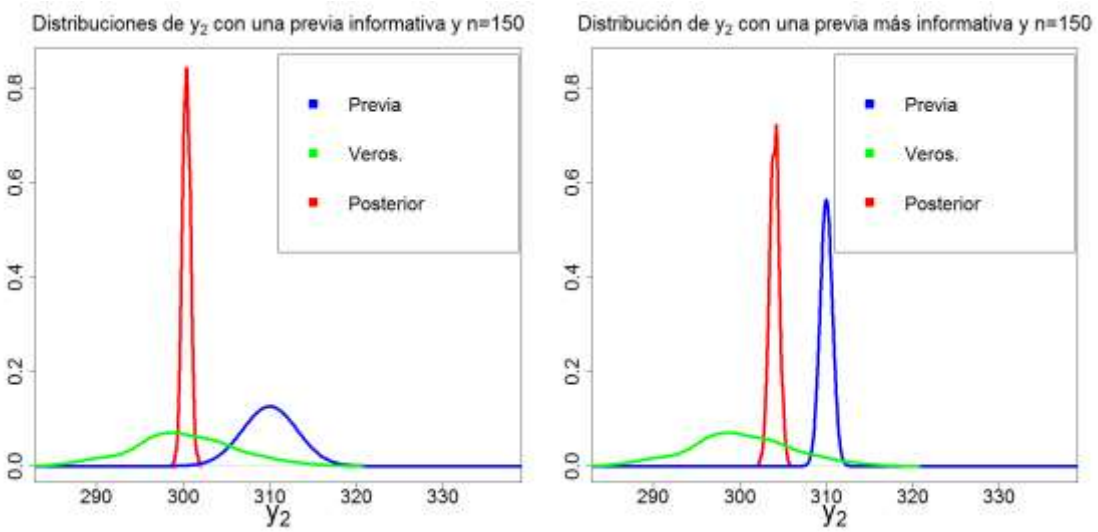
Bestalde, afldez aurretiko lau bat erabiliko bagenu ($\mathrm{mu} \sim \mathrm{N}(0, \frac{1}{1000})$, adibidez), indar guztia egiantzari emango genioke, aldez aurretiko ezagutzarik gabe informazio-iturri bakarra behatutako datuak izango bailirateke. Laginaren tamaina 250 zenean, $\mathrm{mu} \sim \mathrm{N}(30, \frac{1}{0.5})$ aldez aurretikoak aldez aurretiko lau batek bezala jokatzen zuen ia, datu kopuru handia erabiltzen zenez gero aldez aurretikoak ez baitzuen garrantzirik. Grafikoan ikus dezakegunez, a posteriori banaketa eta egiantza balio berean zentratuta daude ia, eta a posteriori batezbestekoaren estimazio puntuala $y_{1.3} = 27.33$ da.



$\boldsymbol{z} = (0,1)$ denean (bigarren estratua), esan dezakegu $Y \sim N(mu, tau)$ eta $mu \sim N(310, \frac{1}{10})$ direla. Aldez aurretiko bariantzari buruzko a priori informaziorik ez dugunez gero, informaziozkoa ez den aldez aurretiko bat esleituko diogu tau-ri: adibidez, $tau \sim Ga(0.001, 0.001)$. Aldez aurretikoari indar handiagoa eman nahi badiogu, bariantza txikiagotu dezakegu. Adibidez, pentsa liteke $\mathrm{mu} \sim \mathrm{N}(310, \frac{1}{0.5})$ dela.

Hurrengo urratsa behatutako datuek ematen diguten informazioa ikustea da. Adibide honetarako, lehen estratuan ditugun 150 behaketak simulatu ditugu, $\mathrm{N}(300, \frac{1}{40})$ banaketari jarraitzen diotela suposatuta, eta 300.14 lagin-batezbestekoa lortu dugu, 32.70 bariantzarekin.

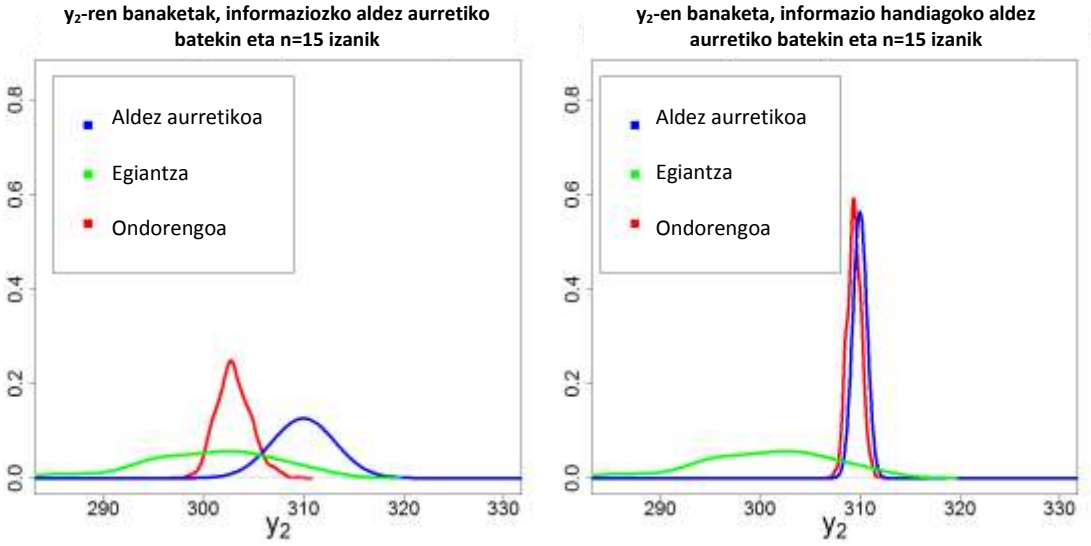
Hirugarren eta azken urratsa a posteriori banaketa lortzea izango da. Berriz ere WinBUGS programa erabiliko dugu, R-tik deituta. Hurrengo grafikoetan, proposatutako bi aldez aurretikoetarako lortutako banaketak agertzen dira:



Estimazio puntual gisa, a posteriori batezbestekoa erabiliko dugu: $\mu_{2.1} = 300.37$ eta $\mu_{2.2} = 303.99$. Arestian aipatu dugun efektu bera gertatzen da, hau da, datu ugarirekin behatzean egiantzak pisu handia duela. Laginaren datuei indarra kentzeko, informazio handiko aldez aurretiko bat beharko litzateke. Era horretan, aldez aurretiko lehenak aldez aurretiko lau batek bezala jokatzen du ia, eta bigarrenak egiantzaren antzeko garrantzia dauka.

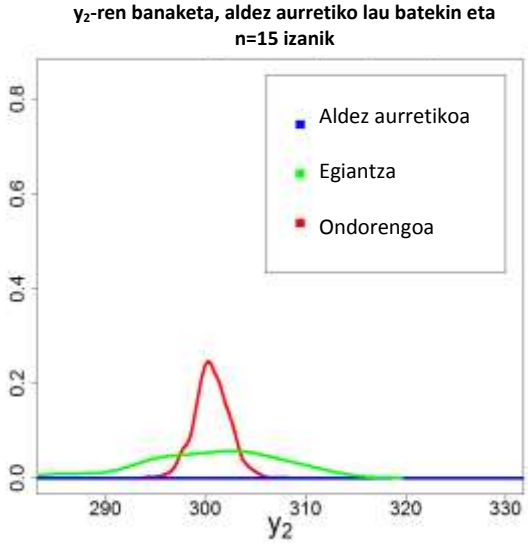
Behatutako datuen lagin-tamainaren eragina hobeto ikusteko, demagun orain laginaren tamaina 150 behaketatik 15era murrizten dugula eta datuak $\mathrm{N}(300, \frac{1}{40})$ banaketa bera erabiliz simulatzen ditugula. Kasu horretan, laginaren batezbestekoa 300.64 izango da, eta bariantza, berriz, 41.42.

Arestian mu, $\mathrm{mu} \sim \mathrm{N}(310, \frac{1}{10})$ eta $\mathrm{mu} \sim \mathrm{N}(310, \frac{1}{0.5})$-rako proposatutako aldez aurretiko biak erabiltzen baditugu, honako grafiko hauek lortuko ditugu:



Berriz ere, a posteriori banaketak eta a posteriori batezbestekoaren estimazio puntualak ($y_{2.1} = 303.04$ eta $y_{2.2} = 309.44$) laginaren tamaina handiagoa zenean lortutakoenen desberdin samarrak dira. Aldez aurretiko lehenarekin, aldez aurretiko batezbestekoaren eta laginaren batezbestekoaren artean dagoen estimazio bat lortzen da. Hau da, kasu horretan, aldez aurretikoak eragina du amaierako a posteriori banaketan. Bigarren kasuan, balio berean zentratuta dagoen, baina bariantza txikiagoa duen aldez aurretiko bat erabiltzen da. Aldez aurretiko eta ondorengo banaketak zertxobait desberdinak dira, ditugun datu urriek (n=15) informazio gutxi ematen baitute.

Azkenik, aldez aurretiko lau bat erabiliko bagenu ($\mathrm{mu} \sim \mathrm{N}(0, \frac{1}{1000})$, adibidez), indar guztia egiantzari emango genioke, aldez aurretiko ezagutzarik gabe informazio-iturri bakarra behatutako datuak izango bailirateke (gutxi izanik ere). Laginaren tamaina 150 zenean, $\mathrm{mu} \sim \mathrm{N}(310, \frac{1}{10})$ aldez aurretikoak aldez aurretiko lau batek bezala jokatzen zuen ia, datu kopuru handia erabiltzen zenez gero aldez aurretikoak ez baitzuen garrantzirik. Grafikoan ikus dezakegunez, a posteriori banaketa eta egiantza balio berean zentratuta daude ia, eta a posteriori batezbestekoaren estimazio puntuala $y_{2.3} = 300.63$ da.



Estratu bakoitzerako emaitzak lortu ondoren, guztizko a posteriori batezbestekoa lortzea izango da helburua. Horretarako, batezbesteko haztatu bat erabiliko da. Lehen eta bigarren estratuen populazio-tamainak 5.000 eta 8.000 unitate dira hurrenez hurren. Horrela, beraz, taulan agertzen diren datuak lortuko dira.

		$\hat{y}_1$	$\hat{y}_2$	$\hat{y}$
n_1= 250 n_2= 150	**Informazio gutxiko aldez aurretikoa**	27.16	300.15	$\frac{5000 * 27.16 + 3000 * 300.15}{8000} = 129.53$
	Informazio handiko aldez aurretikoa	28.86	303.99	$\frac{5000 * 28.86 + 3000 * 303.99}{8000} = 132.03$
n_1= 25 n_2= 15	**Informazio gutxiko aldez aurretikoa**	28.15	303.04	$\frac{5000 * 28.15 + 3000 * 303.04}{8000} = 131.23$
	Informazio handiko aldez aurretikoa	29.88	309.44	$\frac{5000 * 29.88 + 3000 * 309.44}{8000} = 134.72$
	Aldez aurretiko laua	27.36	300.63	$\frac{5000 * 27.36 + 3000 * 300.63}{8000} = 129.84$

Adibide horiekin, agerian geratzen da aldez aurretiko ezagutza askoz garrantzitsuagoa dela laginaren tamaina txikia denean eta informazio gehigarriarekin osatu behar denean. Normalean, aldez aurretiko informazioarekin erabat konbentzituta gaudenean izan ezik, ez da komeni bariantza gehiegi murriztea, horrek eragin negatiboa izan baitezake emaitzetan.

6.2 Adibidea proportzio batekin

Demagun N_j, $j \in \{1,2\}$ tamainako populazio batetik hartutako n_j tamainako ausazko beste bi lagin ditugula. Aurreko kasuan bezala, i behaketa bakoitzari dagokion estratua $\mathbf{z} = (z_1, z_2)$ bektore dikotomikoaren bidez adierazten da, non $z_j = 1$ek adierazten duen behaketa hori *j* estratukoa dela. Azkenik, ezaugarri jakin baten presentzia neurtzen da, non

$$y_i = \begin{cases} 1 \text{ presencia de la variable y en la observación i} \\ 0 \text{ ausencia de la variable y en la observación i} \end{cases}$$

Gure kasuan, alde batetik $n_1 = 50$ eta $n_2 = 30$ $(n = 80)$ dira, eta, bestetik, $N_1 = 500$ eta $N_2 = 300$ $(N = 800)$. Helburua y ezaugarria estratu bakoitzean zer proportziotan dagoen eta guztizkoan zer proportziotan dagoen jakitea da.

Estatistika bayestarraren zailtasunetako bat aldez aurretiko ezagutza probabilitate-banaketa baten bidez adieraztea da. Izan ere, era askotara egin daiteke, eta aukera asko eta asko daude. 2. kapituluan ikusi dugunez, proportzioekin lan egiten denean komeni da beta familiako banaketa bat erabiltzea. Beta banaketa banaketa konjugatua da, hau da, ondorengo banaketak beta banaketa bat izaten jarraitzen du, eta eguneratze-sistema sinplea da. Horrelako banaketa bat a eta b parametroek osatzen dute, eta, ikusi dugunez:

$$\begin{cases} \pi_0 = \dfrac{a}{a+b} \\ \sigma_0 = \sqrt{\dfrac{\pi_0(1-\pi_0)}{(a+b+1)}} \end{cases}.$$

Bestalde, komeni da, halaber, aldez aurretikoaren lagin-tamaina baliokidea kalkulatzea. $\mathrm{Binomial}(n,\pi)$ baten $\hat{\pi} = \frac{y}{n}$ proportzioak honako bariantza hau du: $\frac{\pi(1-\pi)}{n}$. Bariantza hori aldez aurretiko bariantzarekin berdinduz (a priori batezbestekoaren bidez):

$$\frac{\pi_0\,(1-\pi_0)}{n_{eq}} = \frac{ab}{(a+b)^2(a+b+1)}.$$

$\frac{a}{a+b} = \pi_0$ eta $\frac{b}{a+b} = 1-\pi_0$ denez gero, lagin-tamaina baliokidea $n_{eq} = a+b+1$ da. Hau da, aukeratu dugun banaketak ematen duen informazio kantitatea tamaina horretako ausazko lagin baten baliokidea da.

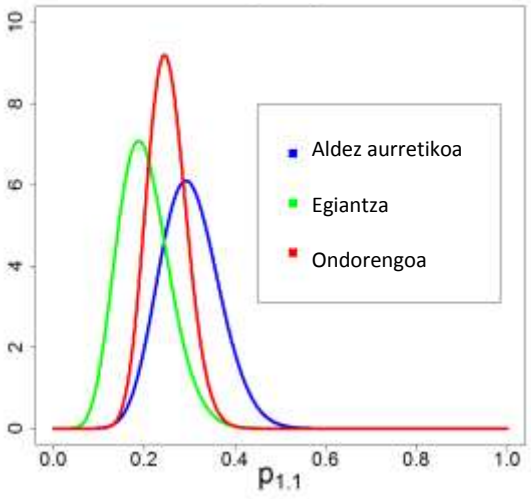
Aurretik egindako azterlan batzuetatik, badugu informazioren bat y-k estratu bakoitzean duen presentziaren inguruan. Demagun $\boldsymbol{z} = (1,0)$ betetzen duten behaketen kasuan (1. estratua) badakigula 50 indibiduoko lagin-tamaina batean oinarrituta arrakastaren proportzioa 0.3 dela. Era horretan, $\mathrm{Beta}(14.7,\ 34.3)$ aldez aurretiko bat lortuko genuke. Proportzioak 0.2 eta 0.4 balioen artean % 95eko konfiantza estatistikoa badu, berriz, beharrezkoak diren ekuazioak bakandu ondoren $\mathrm{Beta}(24.9,58.1)$ bat izango dugu, 84 behaketaren baliokidea den lagin-tamaina batekin. Azkenik, p-ren balioari buruzko aldez aurretiko informaziorik ez bagenu, $\mathrm{Beta}(1,1)$ banaketa bat erabil genezake, banaketa $\mathrm{Uniforme}[0,1]$ baten baliokidea; horrek esan nahiko luke p-ren balio guztiak probabilitate berekoak direla.

Aldez aurretiko banaketa finkatu ondoren, behatutako datuek ematen diguten informazioa ikusi beharko litzateke. Adibide honetarako, lehen estratuko $n_1 = 50$ behaketak simulatu ditugu banaketa $\mathrm{Binomial}(0.25,50)$ bat erabiliz, eta aldagaiaren 10 presentzia eta 40 absentzia lortu ditugu. Hortaz, behatutako probabilitatea p=0,2 da.

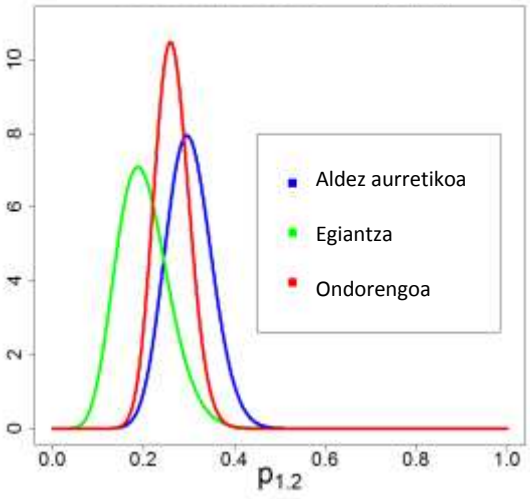
Hirugarren eta azken urratsa a posteriori banaketa lortzea izango da. Aurreko kapituluetan adierazi dugunez, beta banaketa konjugatua da, eta formula sinple baten bidez lortzen da a posteriori banaketa (beta banaketa hura ere). Aldez aurretiko banaketa $\mathrm{Beta}(a,b)$ bat bada eta behatutako datuen arabera $y=10$ eta $n=50$ bada, a posteriori $\mathrm{Beta}(a',b')$ banaketa bat lortuko da, non $a'=a+10$ eta $b'=b+50-10$ den.

Ondorengo grafikoetan, aukeratutako aldez aurretikoaren arabera lortutako banaketak ditugu.

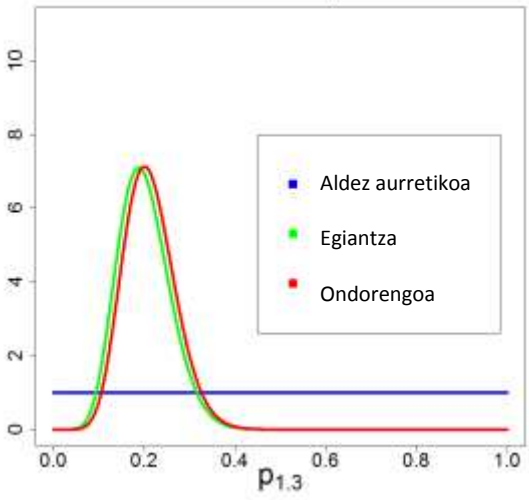
Aldez aurretiko ezagutza: p=0.3 n=50



Aldez aurretiko ezagutza: IC%95=(0.2,0.4)



Aldez aurretiko ezagutzarik gabe



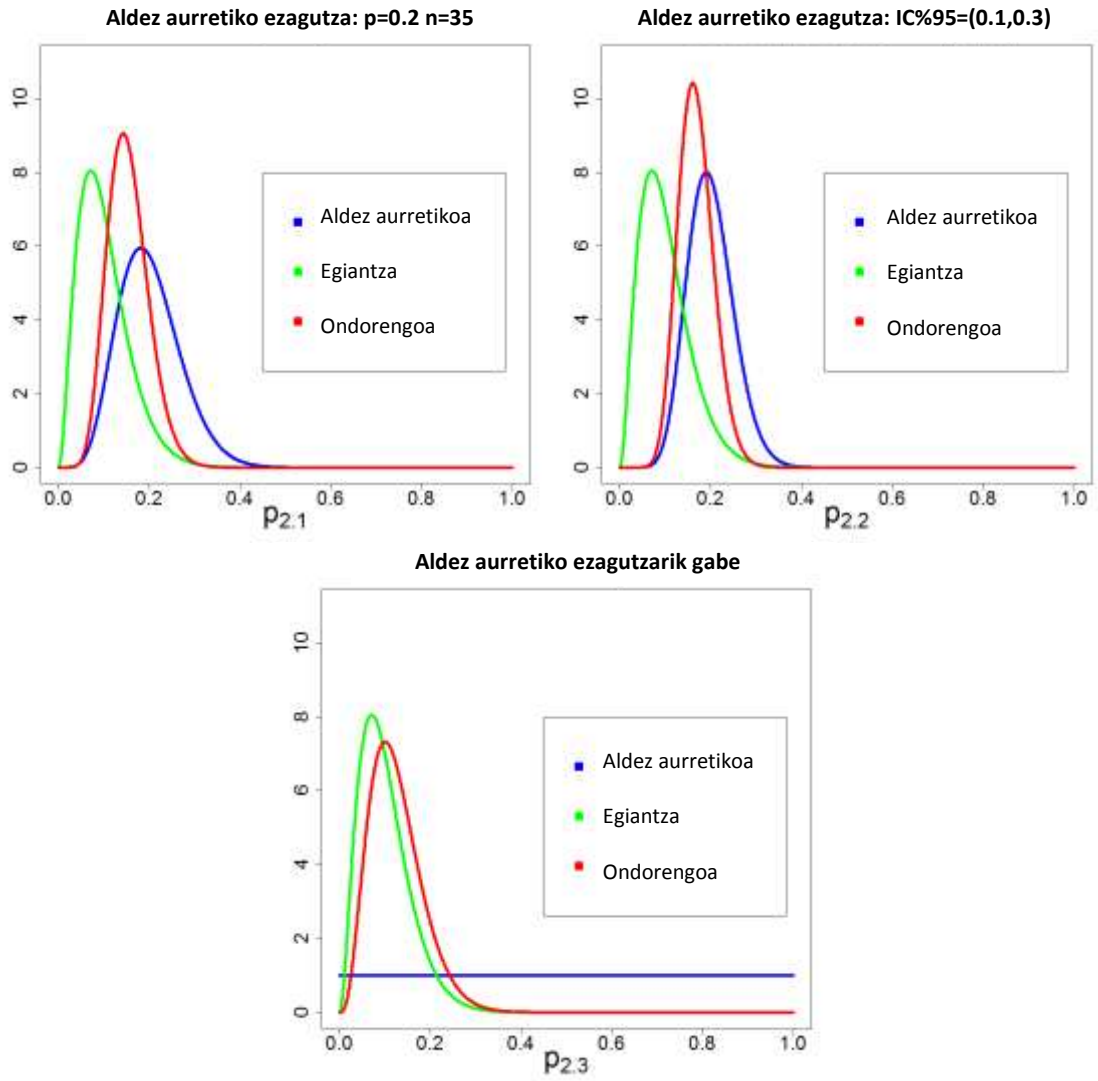
Estimazio puntual gisa, a posteriori batezbestekoa erabiliko dugu:

p1.1=0.2495 p1.2=0.2624 p1.3=0.2115

Lehen bi behaketen artean ez dago hain alde handirik, bi kasuetan antzeko a priori bat baitzegoen. Aldez aurretiko horrek adierazten zuen p-ren balioa 0.3tik hurbil zegoela; egiantzak, aldiz, 0.2tik hurbil zegoela adierazten zigun. Azkenik, bi informazio-iturriak kontuan hartzean, bitarteko balio bat lortzen da. Hala ere, lehen kasuan balioa txikixeagoa da, aldez aurretikoak behaketa kopuru txikiagoa baitzuen. Hirugarren kasuan, berriz, aldez aurretiko ezagutzarik ez genuenez gero, a posteriori batezbestekoa datuetan behatutakoaren oso antzekoa da.

Bigarrenik, demagun $\boldsymbol{z} = (0,1)$ betetzen duten behaketei (2. estratua) buruzko aldez aurretiko informazioa ere badugula. Jakin dezakegu 35 behaketako lagin-tamaina batean oinarrituta proportzioak 0.2 batezbestekoa duela, edota probabilitateak % 95eko konfiantza estatistikoa duela 0.1 eta 0.3 balioen artean. Azkenik, aldez aurretiko informaziorik gabeko kasua dugu. Arestian adierazitako formulak erabiliz, $\mathrm{Beta}(6.8,\ 27.2)$ bat lortzen da lehen kasurako, $\mathrm{Beta}(12.6,\ 50.4)$ bat (63ren tamaina baliokidea) bigarren kasurako, eta aldez aurretiko Beta(1,1) laua hirugarren kasurako.

Behatutako datuak simulatu ditugu, eta 30 behaketatik 3k dute ezaugarria; hau da, p=0.1. Ondoren, eguneratze-formulak erabiliz, irudian ikus ditzakegun a posteriori banaketak kalkulatu ditugu.



Estimazio puntual gisa, a posteriori batezbestekoa erabiliko dugu:

p2.1=0.1531 p2.2=0.1677 p2.3=0.125

Aurreko kasuan bezala, lehen bi a posteriori proportzioak elkarren antzekoak dira, sartutako aldez aurretiko informazioa balio berean zentratuta baitzegoen. Hala ere, ikus daitekeenez, bigarrenak indar handiagoa zuen. Bi balioak aldez aurretiko informazioaren eta behatutakoaren artean daude. Hirugarren kasuan, aldez aurretiko informazio asko ez dagoenez gero, a posteriori proportzioak egiantzaren antzeko balio bat hartu du.
Azkenik, ondorengo taulan ikus daitekeenez, laginketa estratifikatu sinple batekin lan egitean beharrezkoa da lortutako estimazioak haztatzea, proportzioaren estimazio globala lortzeko.

	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}$
Informaziozko aldez aurretikoa, aldez aurretiko ezagutza urriarekin	0.2495	0.1531	$\frac{500 * 0.2495 + 300 * 0.1531}{800} = 0.2134$
Informazio handiko aldez aurretikoa, aldez aurretiko ezagutza handi samarrarekin	0.2624	0.1677	$\frac{500 * 0.2624 + 300 * 0.1677}{800} = 0.2269$
Aldez aurretiko laua, aldez aurretiko ezagutzarik gabe	0.2115	0.125	$\frac{500 * 0.2115 + 300 * 0.125}{800} = 0.1791$

Emaitza globaletan gehiegizko alderik ez badago ere, garbi dago aldez aurretikoaren aukeraketa garrantzitsua dela eta emaitzetan aldaketak eragin ditzakeela.

6.3 Emaitzak

Deskribatutako adibideetan ikusi dugunez, ondorio intuitiboak eta naturalak lortzeko eta aldez aurretiko ikuspuntuak edo usteak gaineratzeko aukera ematen dute metodo bayestarrek.

Bayesen teoremak gure aldez aurretiko usteak datuek ematen duten informazioarekin eguneratzeko aukera ematen digu. Horregatik, datuak aztertu aurretik garrantzitsua da aldez aurretiko ezagutzak behar bezala adierazten jakitea. Komeni da aukeratutako banaketa marraztea, balioak arrazoizkoak direla eta gure ezagutzarekin bat datozela egiaztatzeko. Bat ez badatoz, parametroak alda daitezke kurbak nahi dugun forma eduki arte.

Lan egiteko erabiltzen ditugun datuen bolumena zenbat eta handiagoa izan, orduan eta txikiagoa izango da aldez aurretiko banaketaren eragina. Horrela, beraz, aldez aurretiko banaketa desberdin samar batzuetatik abiatuta ere, oso antzekoak diren a priori banaketak lortu ahal izango ditugu. Ondorio gisa, esan dezakegu inferentzia bayestarrak askoz indar handiagoa duela lagin-tamaina txikiak ditugunean. Izan ere, kasu horietan izaten da beharrezkoa informazio gehigarria, gerta baitaiteke eskuragarri dagoen informazioa urria izatea eta adierazgarria ez izatea.

7. Merkataritza elektronikoaren inguruko eredu frekuentista bat vs. eredu bayestar bat, EUSTATek egindako Informazioaren Gizarteari buruzko Inkesta – Enpresak (IGIE) izenekotik abiatuta

7.1 Testuingurua

Informazioaren Gizarteari buruzko Inkesta – Enpresak (IGIE) 2001ean jarri zuen abian EUSTATek, teknologia berriek EAEko ekonomia-sarean zenbateraino sartu diren zehazteko eta horren inguruko jarraipena egiteko, teknologia-alorrak hartzen ari diren iraultza-izaera dela-eta. Harrez geroztik, urtero egin izan da inkesta.

Laginaren estratuak jarduera ekonomikoaren, enplegu kopuruaren eta Lurralde Historikoaren arabera zehaztu ziren. Hasiera batean, sei enplegu-estratu zituen inkestak: 0-5, 6-9, 10-19, 20-49, 50-99 eta 100 enplegu edo gehiago, eta jarduera ekonomikoak 65 kodetan banatuta zeuden, Ekonomia Jardueren Sailkapen Nazionalaren (EJSN) arabera. Geroago, berriz, jarduera ekonomikoak 27 kategoriatan (2009ko EJSNko A38 sailkapenaren arabera) eta enpleguak 3 estratutan (10 enplegu baino gutxiago, 10-99 enplegu eta 100 enplegu baino gehiago) banatzea erabaki zen.

Beren jarduera EAEn egiten duten edozein tamainatako eta edozein jarduera-sektoretako (lehen sektorea eta etxeko zerbitzua salbuetsita) establezimenduen multzoa da IGIEko erreferentziazko populazioa. Inkesta egin duen populazioa EUSTATen EAEko Jarduera Ekonomikoen Direktoriotik hartu da. Direktorio hori inkesta gehigarri garrantzitsuak egiteko ere erabiltzen da, eta jarduera ekonomiko, Lurralde Historiko, enplegu kopuru eta eskualde bakoitzaren barruan dauden establezimenduen kopuruari buruzko informazioa biltzen du. Saltokien kopurua era irregularrean banatuta dago hiru Lurralde Historikoetako hogei eskualdeetan. Inkesta, nolanahi ere, Lurralde Historikoaren, enplegu kopuruaren eta jarduera ekonomikoaren inguruko zuzeneko zenbatesleak emateko diseinatu zen.

Kapitulu honetan proposatutako aplikaziorako, Gipuzkoan merkataritza elektronikoan diharduten establezimenduen proportzioaren inguruko estimazioetan murgilduko gara. IGIEaren barruan, Gipuzkoako lagina 2.500 establezimendu inguruko panel bat da, % 15-20ko urteko berriztapen-tasarekin. Lehenik eta behin, era bayestarrean zenbatetsitako eredu bat erabiltzea proposatuko dugu merkataritza elektronikoan diharduten establezimenduen proportzioa Lurralde Historikoaren eta jarduera ekonomikoaren arabera finkatzeko. Ondoren, emaitza horiek zuzeneko zenbatesle tradizionalen bidez lortutako emaitzekin konparatuko ditugu.

7.2 Datuak

Gipuzkoaren kasuan eskuragarri ditugun t=2015eko datuak H=169 estratutan eta K=27 jarduera ekonomiko agregatutan egituratuta daude, non

- $\boldsymbol{y_{hk}}$ -k adierazten duen merkataritza elektronikoan diharduten h estratuko eta k jarduera ekonomikoko establezimenduen kopurua.
- $\boldsymbol{n_{hk}}$ -k adierazten duen h estratuan eta k jarduera ekonomikoan dauden lagineko establezimenduen kopurua.
- $\boldsymbol{N_{hk}}$ -k adierazten duen h estratuan eta k jarduera ekonomikoan dauden lagineko establezimenduen kopurua.
- $\boldsymbol{w_{hk}}$ -k adierazten duen h estratuko eta k jarduera ekonomikoko unitate bakoitzaren pisua.

Halaber, aurreko urteetako informazioa dugu (t-1, t-2, ...), era berean egituratuta.

7.3 Metodologia

Gipuzkoan jarduera ekonomiko bakoitzean (k) merkataritza elektronikoan diharduten saltokien proportzioaren estimazioak kalkulatzeko, zuzeneko zenbatesle bat erabiltzen da (zehazki, Horvitz-Thompsonen zenbateslea).

Gipuzkoako proportzio globala zenbatesteko, honako formula hau erabiliko dugu:

$$\hat{\mathrm{p}} = \frac{\sum_{\mathrm{h}} \mathrm{y_h w_h}}{\sum_{\mathrm{h}} \mathrm{w_h}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{h}} \mathrm{N_h} \hat{\mathrm{p}}_{\mathrm{h}} \qquad (7.1).$$

Eta k jarduera ekonomiko bakoitzari dagokion proportzioaren kasuan, berriz,

$$\hat{\mathrm{p}}_{\mathrm{k}} = \left. \frac{\sum_{\mathrm{h}} \mathrm{y_h w_h}}{\sum_{\mathrm{h}} \mathrm{w_h}} \right|_{k} = \left. \frac{1}{\mathrm{N}} \sum_{\mathrm{h}} \mathrm{N_h} \hat{\mathrm{p}}_{\mathrm{h}} \right|_{k}, \quad \forall \, \mathrm{k} \in \mathrm{A38} \qquad (7.2).$$

Non $N = \sum_{\mathrm{h}} \mathrm{N_h}$.

Arazoa orain estratu bakoitzean merkataritza elektronikoan diharduten saltokien proportzioaren estimazioak kalkulatzen data. Hau da, 169 estratu bakoitzarentzat, bere $\hat{\mathrm{p}}_{\mathrm{h}}$ aurkitu behar da. Emaitzak konparatu ahal izateko, probabilitate hauek estatistika frekuentistarekin eta estatistika bayesiarrarekin kalkulatuko ditugu.

7.3.1 Estatistika frekuentista

Estrato bakoitzarentzat, estrato bakoitzean merkataritza elektronikoan diharduten saltokien proportzioa estimatzeko, zatiketa egingo da merkataritza elektronikoan diharduten saltoki eta saltoki guztien kopuruaren artean:

$$\hat{\mathrm{p}}_{\mathrm{h}} = \frac{\mathrm{y_h}}{\mathrm{n_h}}.$$

Ondoren, (7.1) eta (7.2) erabiliz, merkataritza elektronikoan diharduten saltokien proportzioaren estimazioak kalkulatuko dira Gipuzkoa eta jarduera ekonomiko bakoitzerako.

Azkenik, %95ko konfiantza tarteak lortzeko hurrengo formulak erabiliko dira:

$$\mathrm{IC}_{\hat{\mathrm{p}}}^{95\%} = \hat{\mathrm{p}} \pm 1.96 \sqrt{\sum_{\mathrm{h}} \left(\frac{\mathrm{N_h}}{\mathrm{N}}\right)^2 \left(1 - \frac{\mathrm{n_h}}{\mathrm{N_h}}\right) \frac{\hat{\mathrm{p}}_{\mathrm{hk}}(1 - \hat{\mathrm{p}}_{\mathrm{hk}})}{\mathrm{n_{hk}} - 1}}$$

$$IC_{\hat{p}_k}^{95\%} = \hat{p}_k \pm 1.96 \sqrt{\sum_h \left(\frac{N_h}{N_k}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}}, \forall\, k \in A38.$$

Non $N = \sum_h N_h$.

7.3.2 Estatistika bayesiarra

Arestian adierazi dugunez, inferentzia bayestarrak zenbatesleak optimizatzeko aukera ematen du, gure ezagutza eguneratzeko modu bat eskaintzen baitu (behatutako datuak eta aurreko inkestetatik hartutako aldez aurretiko informazio bat, informazio osagarri bat edota adituen iritzia konbinatzea, alegia). Kasu honetan, aurreko urteko IGIEko aldez aurretiko informazioa duen eredu bat erabiliko dugu.

2. kapituluan azaldutako metodologian oinarrituta, estratu bakoitzeko t. urteko erantzun-aldagaiak banaketa $\mathrm{Binomial}$ bati jarraitzen dio, n eta p parametroekin. Era berean, p parametro bakoitzak a eta b parametroak dituen beta banaketa bati jarraitzen dio:

$$y_h{}^t \sim \mathrm{Bin}(n_h{}^t, p_h{}^t)$$

$$p_h{}^t \sim \mathrm{Beta}(a_h, b_h)$$

Estratu bakoitzeko Beta banaketaren a_h eta b_h parametroak kalkulatzeko, honako ekuazio-sistema hau ebatzi beharko dugu:

$$\begin{cases} \hat{p}_h{}^{t-1} = \dfrac{a_h}{(a_h + b_h)} \\ \mathrm{Trunc}\left(\varphi_{h(t,t-1)} * n_h{}^{t-1}\right) = a_h + b_h + 1 \end{cases}$$

Non,

- $\boldsymbol{y_h{}^t}$ t. urtean merkataritza elektronikoan jardun duten saltokien kopurua den.
- $\boldsymbol{n_h{}^t}$ h estratuak t. urtean duen lagin-tamaina den.
- $\boldsymbol{\hat{p}_h{}^t}$ t. urtean h estratu bakoitzean merkataritza elektronikoan aritzeko zenbatetsitako aldez aurretiko probabilitatea den.
- $\boldsymbol{Trunc}(.)$ zati osoa itzultzen digun funtzioa den.
- $\boldsymbol{\varphi_{h(t,t-1)}}$ φ koefizientea edo Mathewsen korrelazio-koefizientea den. Koefiziente hori bi aldagai bitarren arteko elkarketaren neurri bat da. Kasu honetan, t-tik t-1era bitarteko eta h estratu bakoitzaren barruko y erantzun-aldagaiaren elkarketa neurtzen du. Horrela, beraz, korrelazioa zenbat eta handiagoa izan, orduan eta garrantzi handiagoa emango zaio t-1. urteko aldez aurretiko informazioari.

h estratu jakin baterako ϕ koefizientea lortu ezin izan bada, erabili beharreko aldez aurretikoaren tamaina baliokidea 3 unitate izango da. Gainera, estratu jakin bati buruzko aldez aurretiko informaziorik izan ezean, 2. Kapitulua kontuan hartuta $\mathrm{Beta}(1,1)$ banaketa bat erabiliko da, $p=0.5$ baten eta 3 unitateko lagin-tamainaren baliokidea.

Estratu bakoitzean eta t=2015. urtean merkataritza elektronikoan jarduteko a posteriori probabilitateak lortzeko, R-ko R2WinBUGS liburutegia erabili dugu, eta liburutegi horrek WinBUGS software estatistikora jotzen du analisi bayestarra Markov chain Monte Carlo (MCMC) metodoak erabiliz egiteko. Erabilitako funtzio nagusiak dokumentu honen II. Eranskinean agertzen dira.

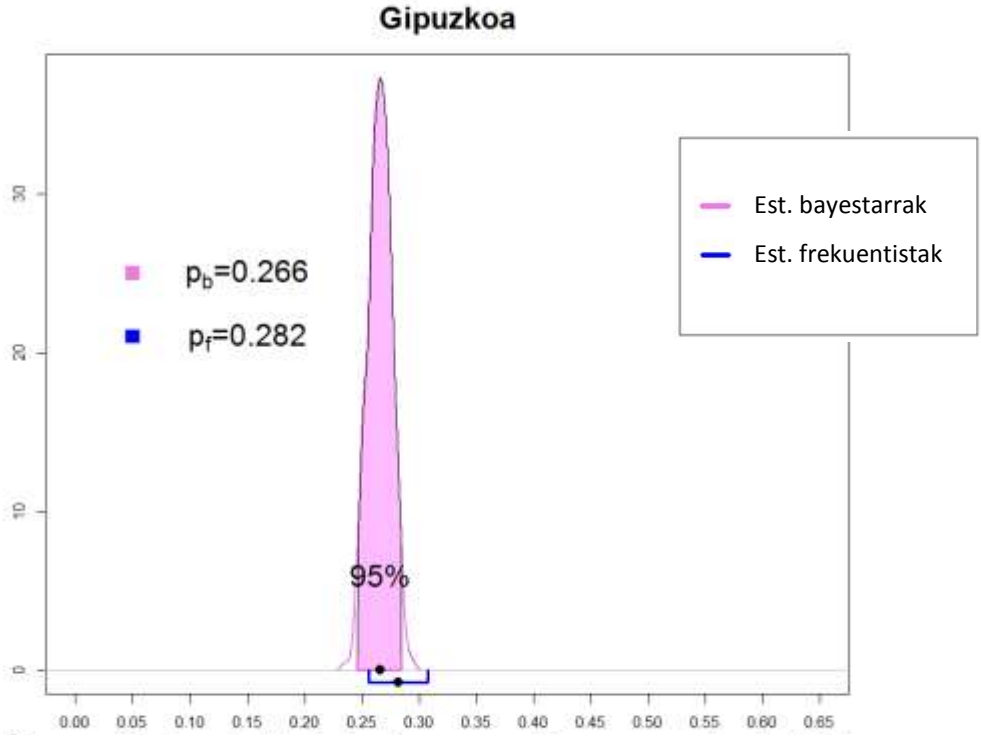
Berriro ere (7.1) eta (7.2) formulak erabiliz, merkataritza elektronikoan diharduten saltokien proportzioaren estimazioak kalkulatuko dira Gipuzkoa eta jarduera ekonomiko bakoitzerako.

7.4 Analisia

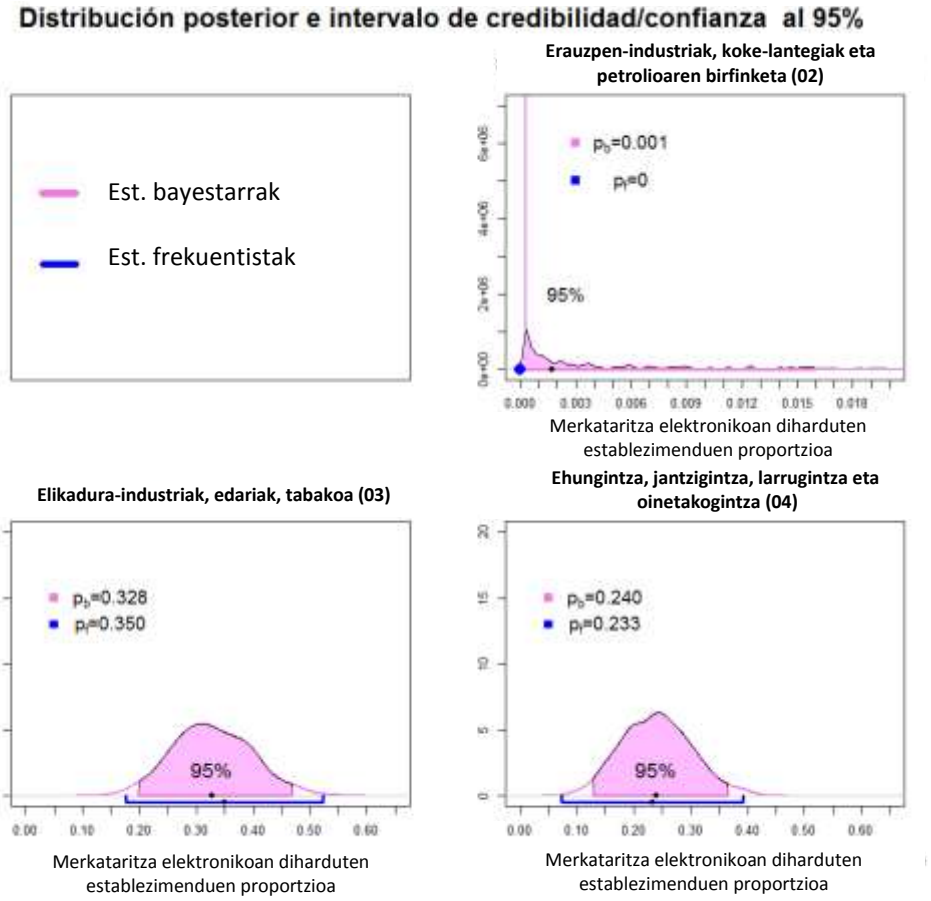
Ondoren, bi paradigmak aplikatzean lortu ditugun emaitzak aurkeztuko ditugu. Lehenik eta behin, Gipuzkoan eta k jarduera ekonomiko bakoitzean merkataritza elektronikoan diharduten saltokien proportzioen inguruko estimazioak (bi metodoei jarraikiz lortutakoak) jaso ditugu ondorengo taulan.

Bigarrenik, bi tekniken bidez lortutako estimazioak eta dagozkien sinesgarritasun-tarteak eta konfiantza-tarteak jasotzen dituzten grafiko batzuk ditugu. Estimazio bayestarren kasuan, parametro bakoitzaren a posteriori dentsitate-funtzioa ere jasota dago.

	Sektorizazioa	Met. bayestarrak		Met. frekuentistak	
		$\hat{p}_b$	Sinesg.-tartea (% 95)	$\hat{p}_f$	Konf.-tartea (% 95)
	Gipuzkoa	0.2660	(0.2465, 0.2845)	0.2819	(0.2565, 0.3074)
A38	Erauzpen-industriak, koke-lantegiak eta petrolioaren birfinketa (02)	0.0017	(0, 0.0160)	0	0
	Elikadura-industriak, edariak, tabakoa (03)	0.3281	(0.1990, 0.4694)	0.35	(0.1772, 0.5228)
	Ehungintza, jantzigintza, larrugintza eta oinetakogintza (04)	0.2401	(0.1281, 0.3654)	0.2329	(0.0727, 0.3931)
	Zura, papera eta arte grafikoak (05)	0.3815	(0.2504, 0.5135)	0.3873	(0.2178, 0.5568)
	Industria kimikoa eta produktu farmazeutikoak (07)	0.3585	(0.2277, 0.5181)	0.3962	(0.2178, 0.5568)
	Kautxua eta plastikoak (09)	0.2173	(0.1321, 0.3167)	0.2440	(0.1191, 0.3688)
	Metalurgia eta produktu metalikoak (10)	0.3001	(0.2131, 0.3987)	0.3355	(0.2070, 0.4640)
	Produktu informatikoak eta elektronikoak. Material eta ekipo elektrikoa (11)	0.4068	(0.3138, 0.5085)	0.3997	(0.2921, 0.5073)
	Makineria eta ekipoa (13)	0.3319	(0.2410, 0.4360)	0.3652	(0.2477, 0.4828)
	Garraio-materiala (14)	0.3460	(0.2207, 0.4831)	0.3628	(0.2057, 0.5199)
	Altzariak eta beste manufaktura batzuk (15)	0.3366	(0.2223, 0.4637)	0.3792	(0.2131, 0.5453)
	Energia elektrikoa, gasa eta lurruna (16)	0.1260	(0.0176, 0.3052)	0.1369	(-0.0186, 0.2924)
	Ur-hornidura eta saneamendua (17)	0.1173	(0.0432, 0.2266)	0.1418	(0.0245, 0.2591)
	Eraikuntza (18)	0.1629	(0.1015, 0.2341)	0.1932	(0.0931, 0.2933)
	Merkataritza eta ibilgailuen konponketa (19)	0.3164	(0.2701, 0.3631)	0.3334	(0.2762, 0.3905)
	Garraioa eta biltegiratzea (20)	0.1379	(0.0809, 0.2127)	0.1327	(0.0455, 0.2199)
	Ostalaritza (21)	0.1499	(0.1096, 0.2014)	0.1649	(0.1031, 0.2268)
	Telekomunikazioak, informatika eta komunikabideak (22)	0.4770	(0.3979, 0.5464)	0.4626	(0.3725, 0.5527)
	Finantza-jarduerak eta aseguruak (25)	0.1811	(0.1174, 0.2538)	0.1815	(0.0994, 0.2635)
	Higiezin-jarduerak (26)	0.2026	(0.1174, 0.3138)	0.2364	(0.1023, 0.3704)
	I+G, aholkularitzak eta beste jarduera profesional eta tekniko batzuk (27)	0.3873	(0.3262, 0.4523)	0.3891	(0.3134, 0.4648)
	Zerbitzu osagarriak (29)	0.2225	(0.1350, 0.3287)	0.2491	(0.1243, 0.3740)
	Bidaia-agentziak eta bidaia-agentzia handizkariak (30)	0.3745	(0.2335, 0.5237)	0.3884	(0.2118, 0.5650)
	Administrazio Publikoa eta defentsa (31)	0.1788	(0.1006, 0.2836)	0.2014	(0.0779, 0.3249)
	Hezkuntza (32)	0.3582	(0.2810, 0.4493)	0.3654	(0.2478, 0.4829)
	Osasun-jarduerak (33)	0.3428	(0.2383, 0.4556)	0.3862	(0.2455, 0.5270)
	Gizarte-jarduerak, jolas-jarduerak, kultura-jarduerak eta beste batzuk (34)	0.1928	(0.1515, 0.2396)	0.2027	(0.1462, 0.2592)

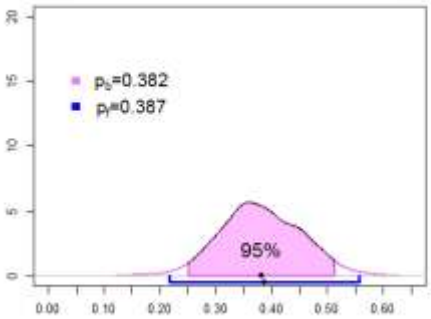


Merkataritza elektronikoan diharduten establezimenduen proportzioa



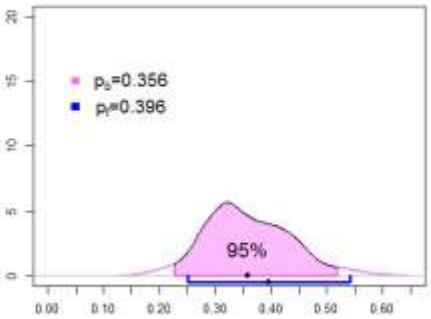
Aktibitate sektorizazioa A38

Zura, papera eta arte grafikoak (05)



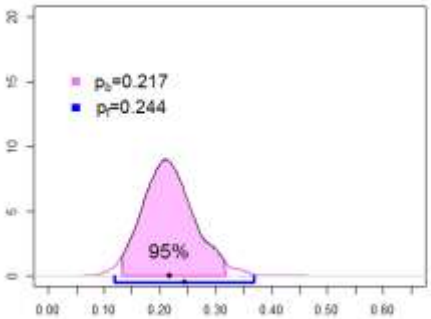
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Industria kimikoa eta produktu farmazeutikoak (07)



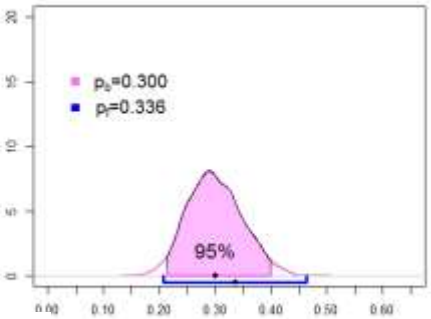
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Kautxua eta plastikoak (09)



Merkataritza elektronikoan diharduten establezimenduen proportzioa

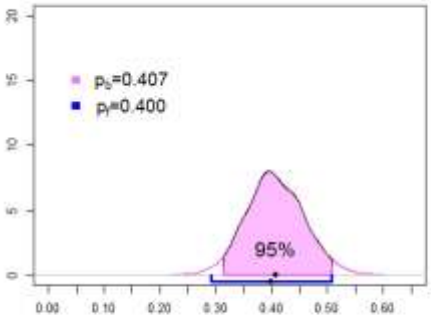
Metalurgia eta produktu metalikoak (10)



Merkataritza elektronikoan diharduten establezimenduen proportzioa

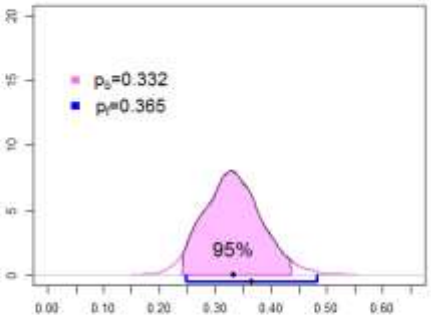
Aktibitate sektorizazioa A38

Produktu informatikoak eta elektronikoak. Material eta ekipo elektrikoa (11)



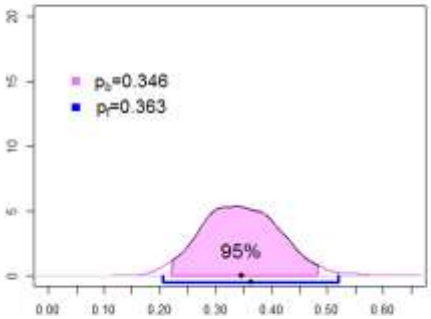
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Makineria eta ekipoa (13)



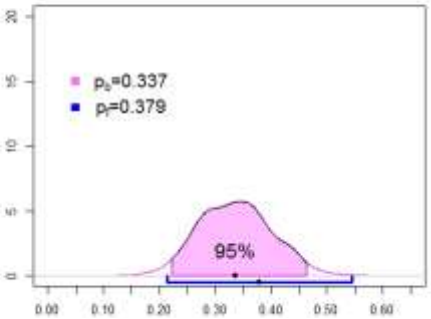
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Garraio-materiala (14)



Merkataritza elektronikoan diharduten establezimenduen proportzioa

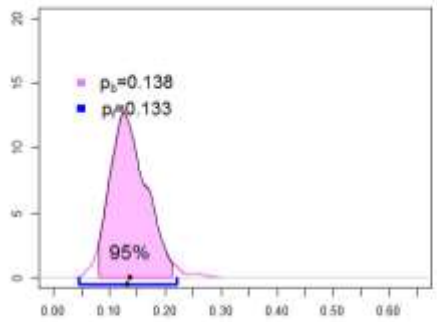
Altzariak eta beste manufaktura batzuk (15)



Merkataritza elektronikoan diharduten establezimenduen proportzioa

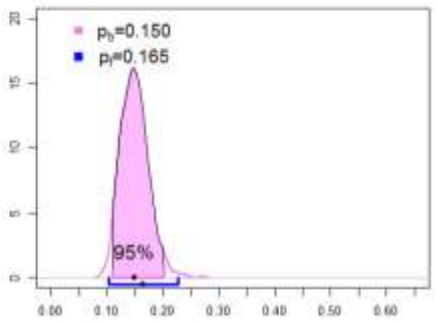
Aktibitate sektorizazioa A38

Garraioa eta biltegiratzea (20)



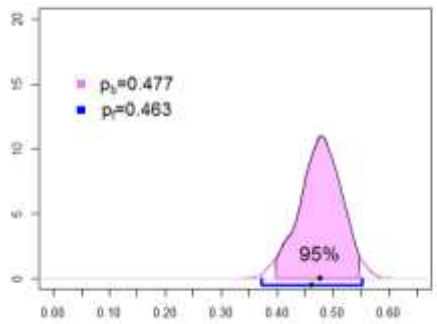
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Ostalaritza (21)



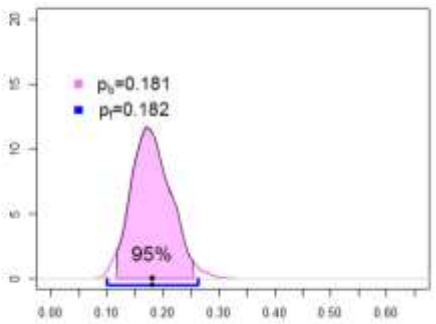
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Telekomunikazioak, informatika eta komunikabideak (22)



Merkataritza elektronikoan diharduten establezimenduen proportzioa

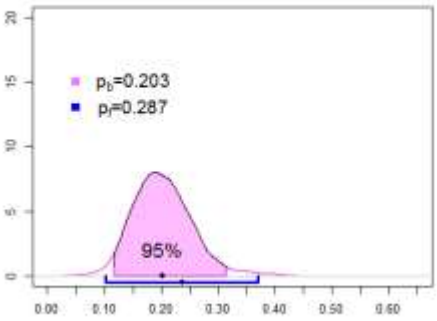
Finantza-jarduerak eta aseguruak (25)



Merkataritza elektronikoan diharduten establezimenduen proportzioa

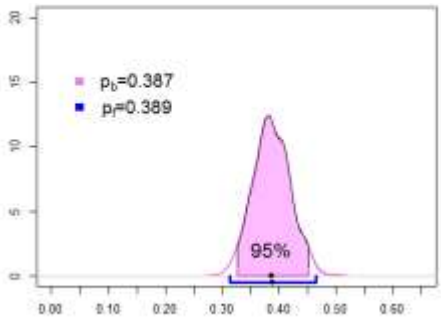
Aktibitate sektorizazioa A38

Higiezin-jarduerak (26)



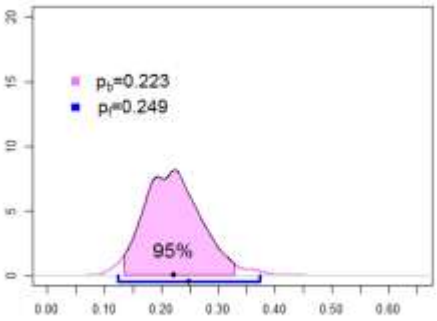
Merkataritza elektronikoan diharduten establezimenduen proportzioa

I+G, aholkularitzak eta beste jarduera profesional eta tekniko batzuk (27)



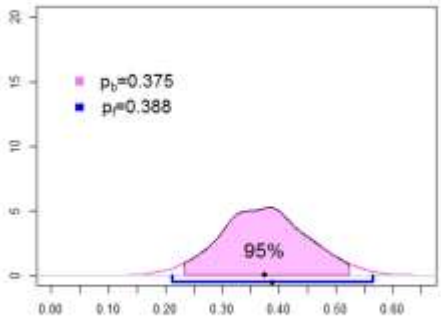
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Zerbitzu osagarriak (29)



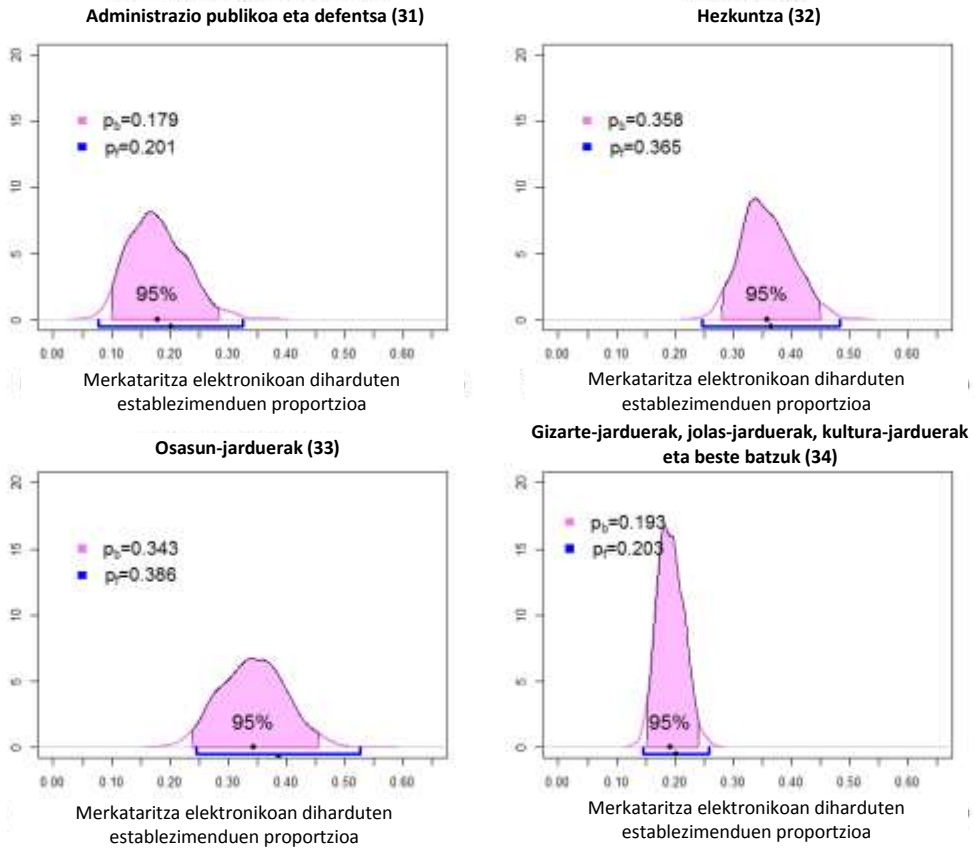
Merkataritza elektronikoan diharduten establezimenduen proportzioa

Bidaia-agentziak eta bidaia-agentzia handizkariak (30)



Merkataritza elektronikoan diharduten establezimenduen proportzioa

Aktibitate sektorizazioa A38



Aktibitate sektorizazioa A38

7.5 Emaitzak

Eredu bayestarrarekin lortutako zenbatesleak, oro har, metodo tradizionalekin lortutakoak baino efizienteagoak izaten dira. Teknika bayestarren bidez, tarteen zabalera murritz daiteke eta denbora-korrelazioari euts dakioke. Gainera, ausazko aldagai bakoitzaren a posteriori probabilitatearen banaketa lor daiteke, eta hipotesiak kontrastatzeko aukera ematen du horrek. Hau da, balio horrek balio jakin bat baino txikiagoa edo handiagoa izateko duen probabilitatea ezar dezakegu estimazio bakoitzean.

Zuzeneko zenbatesleek baliozkoak eta/edo efizienteak ez diren emaitzak eman ditzakete, laginaren tamaina mugatua denean edo behaketa gehienak aldagaiak hartzen dituen balioen muturretako batean daudenean. Taulan, bi datu bitxi ikus ditzakegu. Lehenik eta behin, "Erauzpen-industriak, koke-lantegiak eta petrolioaren birfinketa (02)" jarduera ekonomikoaren kasuan, zenbatetsitako proportzioa 0 da, konfiantza-tartea bezalaxe. Izan ere, laginean jarduera honi lotuta sartu diren saltoki urriak ez dira merkataritza elektronikoan aritzen, eta, beraz, eskuragarri dauden datuek emaitza hori dakarte. Bigarrenik, "Energia elektrikoa, gasa eta lurruna (16)" jardueraren konfiantza-tartearen beheko muturra negatiboa da. Ez du zentzurik probabilitate bat 0 baino txikiagoa izateak, eta, beraz, balio hori 0 batekin zuzendu beharko da eskuz. Eredu bayestarraren bidez, arazo horiek zuzenduta geratzen dira. Izan ere, ereduak eskuragarri dagoen informazio osagarri eta aldez aurretiko guztia baliatzeko eta laginaren tamaina baliokidea handitzeko aukera ematen du, eta emaitzak koherenteak izaten dira beti.

8. Ondorioak

Azkenik, Estatistika eta Matematika Metodologietan prestatzeko eta ikertzeko Bekari esker egindako lan honen ondorio nagusiak azalduko ditugu kapitulu honetan.

8.1 Paradigma

Ikuspegi frekuentistaren eta ikuspegi bayestarraren artean dagoen alde nagusia zera da, lehenak datuak eskuratu aurretik zegoen informazio guztia –teorikoa zein enpirikoa– baztertzen duela. Estatistika bayestarrean, ordea, ez da hutsetik abiatuta lan egiten. Aitzitik, aldez aurretiko ezagutza formalki eta esplizituki hartzen da kontuan, eta probabilitate subjektiboak biltzen ditu, hau da, ikertzailearen intuiziotik sortzen direnak. Probabilitate subjektibo horiek ez dute zertan arrazoi intuitibo soiletan oinarrituta egon. Aitzitik, zerikusi handia dute ikertzailearen trebeziarekin eta aldez aurretiko esperientziarekin.

Estatistika bayestarra baliagarria izan daiteke laginaren tamaina nabarmen murrizten denean. Izan ere, aldez aurretiko datu orokorrez gain gertaera berriak eta ikertzaileek berek errealitatearen portaeraren inguruan dituzten aurreikuspenak kontuan hartzen direnez gero, datuak ez dira indargabetzen. Hau da, estatistika bayestarraren abantaila nagusia isolatuta ez diharduela da. Izan ere, aldez aurretiko ezagutzan oinarrituta dago, eta eskaintzen duen analisi-esparrua askoz aberatsagoa da, ikertzaileari askatasuna ematen baitio bere esperimentuetan emaitza jakin batzuk lortzeko probabilitatea baloratzeko.

Ikuspegi bayestarraren ezaugarri bereizgarri bat zera da, laginaren tamaina oso handia denean, aukeratzen den a priori banaketak oso garrantzi txikia izan ohi duela eta egiantzaren mende dagoela. Zirkunstantzia horrek agerian uzten du metodo bayestarra batez ere laginaren tamainak oso handiak ez direnean dela baliagarria.

8.2 Konputazioa

Teknika bayestarrak inplementatzeko, konputazio-ahalegin handia egin behar izaten da. Zenbait kasutan, beharrezkoak diren integral anizkoitzak ezin dira analitikoki ebatzi. Horregatik, integralak hainbat dimentsiotan kalkulatzea edo hurbiltzea ahalbidetuko duten zenbakizko metodo efizienteak behar dira. Zorionez, XX. mendearen bigarren erdialdean simulazio estokastikoko metodoetan oinarritutako zenbakizko teknika malgu eta efizienteak garatzen hasi ziren. Horren eta birtualizazioari esker konputu-ekipoen prozesatze- eta biltegiratze-ahalmen handiagoa ekarri duen teknologiaren garapenaren ondorioz, metodo bayestarrak gora egiten ari dira.

Gaur egun, oso ahalmen handiko softwarea eta metodologiak daude edonoren eskueran. Lan honetan, nagusiki R eta WinBUGS (Bayesian Inference Using Gibbs Sampling) erabili ditugu estimazio bayestarrak egiteko, baina aukera horiek merkatuan eskuragarri dauden aukera ugarietako bi baino ez dira.

8.3 Estatistika ofiziala

Teknika berriak –eta teknika bayestarrak bereziki– erabiltzeko interesa gero eta handiagoa da Estatistika Ofizialaren barruan. Gero eta ohikoagoa da erakundeei eremu txikien edo hasieran planifikatu ez ziren eta, horregatik, laginean ongi ordezkatuta ez dauden (edota, besterik gabe, ordezkatuta ez dauden) eremu batzuen inguruko estimazioak egin ditzaten eskatzea.

Normalean, eskualde handietarako edo eremu txikietako agregatuetarako estimazio fidagarriak emango dituen lagin-tamaina bat aukeratzen da, baina eremu handietarako erabilitako estatistika-metodoak nekez aplika daitezke eremu txikietan. Izan ere, zuzeneko zenbatesle klasikoek oso errore estandar handiak izaten dituzte, edota ezin dira kalkulatu, planifikatu gabeko eremuren bateko laginik ez dagoelako.

7. Kapituluan ikusi dugunez, planteamendu bayestar alternatibo bat estratu bakoitzari buruz eskuragarri dagoen informaziotik abiatuta aldez aurretiko banaketa bat sortzea eta, era horretan, laginaren datuek ematen duten ezagutza handiagoa edukitzea da. Aurkeztutako aplikazioan, inkesta beretik eta eremu txiki berean lortutako aurreko urteko informazioa erabiltzen zen, denbora-korrelazio berari jarraitzen zitzaion eta estimazio efizienteagoak lortzen ziren.

Beste planteamendu orokorrago bat eredu misto hierarkizatuena da. Eredu horietan, eskuragarri dagoen aldez aurretiko informazioa zein informazio osagarria erabiltzen da eredu aurresalea sortzeko. Hau da, datuen hierarkiak, lotutako eremuek, aurreko urteetako informazioak edota horiek guztiek ematen duten informazioa erabiltzen da eredu beraren barruan. Ereduetan oinarritutako zenbatesleak teknika bayestarren edo teknika frekuentisten bidez lor daitezke.

I. Eranskina

```
#Estimación de una media normal

#Primer estrato

#Simulación de datos N(27,5), n=250

set.seed(1)
n1<-250
y1<-rnorm(n1,27,sqrt(5))

#Resumen
mean(y1)# 27.04957
var(y1)# 4.63241

#Verosimilitud
plot(density(y1))

data.m1<-list(y1=y1,n1=n1)
par.m1 <- c("mu1","var1","sd1")

inits<-function(){
  list(mu1=runif(1,25,35),tau1=runif(1,0.00001,0.01))}

#Cargamos librería
library(R2WinBUGS)

#Previa 1, N(30,0.5)
mu_p1.1<-30
var_p1.1<-0.5
bmbugs1 <- bugs(data.m1,inits, par.m1, "M1.txt",
                n.chains = 1, n.iter = 10000, n.burnin=1000,
                bugs.directory ="...",debug=F)

print(bmbugs1,digits=4)
#27.1545

mu1<- bmbugs1$sims.array[,1 ,"mu1"]
quantile(mu1, c(0.025, 0.975))
sd1<- bmbugs1$sims.array[,1 ,"sd1"]
quantile(sd1, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu1),  main="",cex.axis=2,  xlab="",lwd=5,col="red",  ylim=c(0,3),
xlim=c(18,45),ylab="")
title(main=expression(paste("Distribuciones de ", mu[1]," con una previa
informativa y n=250")),cex.main=2.1,xlab=expression(mu[1]),cex.lab=3)
lines(z,dnorm(z,mu_p1.1,sqrt(var_p1.1)), main="",col="blue",type="l",lwd = 5)
lines(density(y1),ylab="",col="green",type="l",lwd = 5)
legend(31,2.5,nombres,pch=c(15,15,15),col=colores,cex=2)
```

```
#Previa 2, N(30,0.02)
mu_p1_2<-30
var_p1_2<-0.02
bmbugs1_2 <- bugs(data.m1,inits, par.m1, "M1_2.txt",n.chains = 1, n.iter = 10000,
n.burnin=1000, bugs.directory ="...",debug=F)

print(bmbugs1_2,digits=4)
#28.8602

mu1_2<- bmbugs1_2$sims.array[,1 ,"mu1"]
quantile(mu1_2, c(0.025, 0.975))
sd1_2<- bmbugs1_2$sims.array[,1 ,"sd1"]
quantile(sd1_2, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu1_2),  main="",cex.axis=2,  xlab="",lwd=5,col="red",  ylim=c(0,3),
xlim=c(18,45),ylab="")
title(main=expression(paste("Distribuciones de ", mu[1]," con una previa más
informativa y n=250")),cex.main=2.1,xlab=expression(mu[1]),cex.lab=3)
lines(z,dnorm(z,mu_p1_2,sqrt(var_p1_2)), main="",col="blue",type="l",lwd = 5)
lines(density(y1),ylab="",col="green",type="l",lwd = 5)
legend(31,2.5,nombres,pch=c(15,15,15),col=colores,cex=2)

#Disminuímos tamaño muestral del primer estrato

#Simulación de datos, N(27,5) n=25
set.seed(1)
n1_n<-25
y1_n<-rnorm(n1_n,27,sqrt(5))

#Resumen
mean(y1_n)# 27.3772
var(y1_n)# 4.5135

#Verosimilitud
plot(density(y1_n))

data.m1_n<-list(y1=y1_n,n1=n1_n)
par.m1 <- c("mu1","var1","sd1")

inits<-function(){
  list(mu1=runif(1,25,35),tau1=runif(1,0.00001,0.01))}

#Previa 1, N(30,0.5)
mu_p1.1<-30
var_p1.1<-0.5
bmbugs1_n <- bugs(data.m1_n,inits, par.m1, "M1.txt",
                  n.chains = 1, n.iter = 10000, n.burnin=1000,
                  bugs.directory ="...",debug=F)

print(bmbugs1_n,digits=4)
#28.1486
```

```
mu1_n<- bmbugs1_n$sims.array[,1 ,"mu1"]
quantile(mu1_n, c(0.025, 0.975))
sd1_n<- bmbugs1_n$sims.array[,1 ,"sd1"]
quantile(sd1_n, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu1_n),main="",cex.axis=2,   xlab="",lwd=5,col="red",   ylim=c(0,3),
xlim=c(18,45),ylab="")
title(main=expression(paste("Distribución   de   ",   mu[1],"   con   una   previa
informativa y n=25")),cex.main=2.1,xlab=expression(mu[1]),cex.lab=3)
lines(z,dnorm(z,mu_p1.1,sqrt(var_p1.1)), main="",col="blue",type="l",lwd = 5)
lines(density(y1),ylab="",col="green",type="l",lwd = 5)
legend(30,2.5,nombres,pch=c(15,15,15),col=colores,cex=2)

#Previa 2, N(30,0.02)
mu_p1_2<-30
var_p1_2<-0.02
bmbugs1_2_n <- bugs(data.m1_n,inits, par.m1, "M1_2.txt",n.chains = 1, n.iter =
10000, n.burnin=1000, bugs.directory ="…",debug=F)

print(bmbugs1_2_n,digits=4)
#29.8761

mu1_2_n<- bmbugs1_2_n$sims.array[,1 ,"mu1"]
quantile(mu1_2_n, c(0.025, 0.975))
sd1_2_n<- bmbugs1_2_n$sims.array[,1 ,"sd1"]
quantile(sd1_2_n, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu1_2_n), main="",cex.axis=2, xlab="",lwd=5,col="red", ylim=c(0,3),
xlim=c(18,45),ylab="")
title(main=expression(paste("Distribución de ", mu[1]," con una previa más
informativa y n=25")),cex.main=2.1,xlab=expression(mu[1]),cex.lab=3)
lines(z,dnorm(z,mu_p1_2,sqrt(var_p1_2)), main="",col="blue",type="l",lwd = 5)
lines(density(y1),ylab="",col="green",type="l",lwd = 5)
legend(31.5,2.5,nombres,pch=c(15,15,15),col=colores,cex=2)

#Previa 3, N(0,1000)
mu_p1_3<-0
var_p1_3<-1000
bmbugs1_3_n <- bugs(data.m1_n,inits, par.m1, "M1_plana.txt",n.chains = 1, n.iter
= 10000, n.burnin=1000, bugs.directory ="...",debug=F)

print(bmbugs1_3_n,digits=4)
#27.3572

mu1_3_n<- bmbugs1_3_n$sims.array[,1 ,"mu1"]
quantile(mu1_3_n, c(0.025, 0.975))
```

```
sd1_3_n<- bmbugs1_3_n$sims.array[,1 ,"sd1"]
quantile(sd1_3_n, c(0.025, 0.975))

#Gráficos
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu1_3_n),main="",cex.axis=2,  xlab="",lwd=5,col="red",  ylim=c(0,3),
xlim=c(18,45),ylab="")
title(main=expression(paste("Distribución de ", mu[1]," con una previa plana y
n=25")),cex.main=2.1,xlab=expression(mu[1]),cex.lab=3)
lines(z,dnorm(z,mu_p1_3,sqrt(var_p1_3)), main="",col="blue",type="l",lwd = 5)
lines(density(y1),ylab="",col="green",type="l",lwd = 5)
legend(30,2.5,nombres,pch=c(15,15,15),col=colores,cex=2)

##Segundo estrato

#Simulación de datos N(300,40), n=150
set.seed(1)
n2<-150
y2<-rnorm(n2,300,sqrt(40))

#Resumen
mean(y2) #300.1376
var(y2)#   32.70165

#Verosimilitud
plot(density(y2))

data.m2<-list(y2=y2,n2=n2)
par.m2 <- c("mu2","var2","sd2")

inits<-function(){
  list(mu2=runif(1,350,450),tau2=runif(1,0.00001,0.01))}

mu_p2.1<-310
var_p2.1<-10

bmbugs2 <- bugs(data.m2,inits, par.m2, "M2.txt",
                n.chains = 1, n.iter = 10000, n.burnin=1000,
                bugs.directory ="D:/DATOS/NGUTIEAR/Desktop/WinBUGS14/",debug=F)

print(bmbugs2,digits=4)
#300.3694

mu2<- bmbugs2$sims.array[,1 ,"mu2"]
quantile(mu2, c(0.025, 0.975))
sd2<- bmbugs2$sims.array[,1 ,"sd2"]
quantile(sd2, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")
```

```
plot(density(mu2),          main="",cex.axis=2,          xlab="",lwd=5,col="red",
xlim=c(285,337),ylim=c(0,0.85), ylab="")
title(main=expression(paste("Distribuciones de ", mu[2]," con una previa
informativa y n=150")),cex.main=2.1,xlab=expression(mu[2]),cex.lab=3)
lines(z,dnorm(z,mu_p2.1,sqrt(var_p2.1)), main="",col="blue",type="l",lwd = 5)
lines(density(y2),ylab="",col="green",type="l",lwd = 5)
legend(311,0.87,nombres,pch=c(15,15,15),col=colores,cex=2)

#Previa 2, N(310,0.5)
mu_p2.2<-310
var_p2.2<-0.5

bmbugs2_2 <- bugs(data.m2,inits, par.m2, "M2_2.txt",n.chains = 1, n.iter = 10000,
n.burnin=1000, bugs.directory ="...",debug=F)

print(bmbugs2_2,digits=4)
#303.9940

mu2_2<- bmbugs2_2$sims.array[,1 ,"mu2"]
quantile(mu2_2, c(0.025, 0.975))
sd2_2<- bmbugs2_2$sims.array[,1 ,"sd2"]
quantile(sd2_2, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu2_2),         main="",cex.axis=2,         xlab="",lwd=5,col="red",
xlim=c(285,337),ylim=c(0,0.85), ylab="")
title(main=expression(paste("Distribución de ", mu[2]," con una previa más
informativa y n=150")),cex.main=2.1,xlab=expression(mu[2]),cex.lab=3)
lines(z,dnorm(z,mu_p2.2,sqrt(var_p2.2)), main="",col="blue",type="l",lwd = 5)
lines(density(y2),ylab="",col="green",type="l",lwd = 5)
legend(311,0.87,nombres,pch=c(15,15,15),col=colores,cex=2)

#Disminuimos tamaño muestral

#Simulación de datos N(300,40), n=15
set.seed(1)
n2_n<-15
y2_n<-rnorm(n2_n,300,sqrt(40))

#Resumen
mean(y2_n) #300.6378
var(y2_n)#   41.42809

#Verosimilitud
plot(density(y2_n))

data.m2_n<-list(y2=y2_n,n2=n2_n)
par.m2 <- c("mu2","var2","sd2")

inits<-function(){
  list(mu2=runif(1,350,450),tau2=runif(1,0.00001,0.01))}
```

```
#Previa 1, N(310, 10)
mu_p2.1<-310
var_p2.1<-10

bmbugs2_n <- bugs(data.m2_n,inits, par.m2, "M2.txt",
                  n.chains = 1, n.iter = 10000, n.burnin=1000,
                  bugs.directory ="…",debug=F)

print(bmbugs2_n,digits=4)
#303.0441

mu2_n<- bmbugs2_n$sims.array[,1 ,"mu2"]
quantile(mu2_n, c(0.025, 0.975))
sd2_n<- bmbugs2_n$sims.array[,1 ,"sd2"]
quantile(sd2, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu2_n),main="",cex.axis=2,                  xlab="",lwd=5,col="red",
xlim=c(285,330),ylim=c(0,0.85), ylab="")
title(main=expression(paste("Distribución de ", mu[2]," con una previa
informativa y n=15")),cex.main=2.1,xlab=expression(mu[2]),cex.lab=3)
lines(z,dnorm(z,mu_p2.1,sqrt(var_p2.1)), main="",col="blue",type="l",lwd = 5)
lines(density(y2_n),ylab="",col="green",type="l",lwd = 5)
legend(285,0.85,nombres,pch=c(15,15,15),col=colores,cex=2)

#Previa 2, N(310, 0.5)
mu_p2.2<-310
var_p2.2<-0.5

bmbugs2_2_n <- bugs(data.m2_n,inits, par.m2, "M2_2.txt",n.chains = 1, n.iter =
10000, n.burnin=1000, bugs.directory ="...",debug=F)

print(bmbugs2_2_n,digits=4)
#309.4350

mu2_2_n<- bmbugs2_2_n$sims.array[,1 ,"mu2"]
quantile(mu2_2, c(0.025, 0.975))
sd2_2_n<- bmbugs2_2_n$sims.array[,1 ,"sd2"]
quantile("sd2_2_n", c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu2_2_n),        main="",cex.axis=2,        xlab="",lwd=5,col="red",
xlim=c(285,330),ylim=c(0,0.85), ylab="")
title(main=expression(paste("Distribución de ", mu[2]," con una previa más
informativa y n=15")),cex.main=2.1,xlab=expression(mu[2]),cex.lab=3)
lines(z,dnorm(z,mu_p2.2,sqrt(var_p2.2)), main="",col="blue",type="l",lwd = 5)
lines(density(y2_n),ylab="",col="green",type="l",lwd = 5)
legend(284,0.85,nombres,pch=c(15,15,15),col=colores,cex=2)
```

```
#Previa plana, N(0,1000)
mu_p2_3<-0
var_p2_3<-1000
bmbugs2_3_n <- bugs(data.m2_n,inits, par.m2, "M2_plana.txt",n.chains = 1, n.iter
= 10000, n.burnin=1000, bugs.directory ="...",debug=F)

print(bmbugs2_3_n,digits=4)
#300.6254

mu2_3_n<- bmbugs2_3_n$sims.array[,1 ,"mu2"]
quantile(mu2_3_n, c(0.025, 0.975))
sd2_3_n<- bmbugs2_3_n$sims.array[,1 ,"sd2"]
quantile(sd2_3_n, c(0.025, 0.975))

#Gráfico
z <- seq(0,1000,length=10000)

nombres<-list("Previa","Veros.","Posterior")
colores<-c("blue","green","red")

plot(density(mu2_3_n),        main="",cex.axis=2,        xlab="",lwd=5,col="red",
ylim=c(0,0.85), xlim=c(285,330),ylab="")
title(main=expression(paste("Distribución de ", mu[2]," con una previa plana y
n=15")),cex.main=2.1,xlab=expression(mu[2]),cex.lab=3)
lines(z,dnorm(z,mu_p2_3,sqrt(var_p2_3)), main="",col="blue",type="l",lwd = 5)
lines(density(y2_n),ylab="",col="green",type="l",lwd = 5)
legend(306,0.85,nombres,pch=c(15,15,15),col=colores,cex=2)


#M1.txt
model{
#Likelihood
for (i in 1:n1){
y1[i]~dnorm(mu1,tau1)}
#Prior
mu1~dnorm(30,2)
tau1~dgamma(0.0001,0.0001)
var1<-1/tau1
sd1<-sqrt(var1)}

#M1_2.txt
model{
#Likelihood
for (i in 1:n1){
y1[i]~dnorm(mu1,tau1)}
#Prior
mu1~dnorm(30,50)
tau1~dgamma(0.0001,0.0001)
var1<-1/tau1
sd1<-sqrt(var1)}

#M1_plana.txt
model{
#Likelihood
for (i in 1:n1){
y1[i]~dnorm(mu1,tau1)}
#Prior
mu1~dnorm(30,2)
```

```
tau1~dgamma(0.0001,0.0001)
var1<-1/tau1
sd1<-sqrt(var1)}

#M2.txt
model{
#Likelihood
for (i in 1:n2){
y2[i]~dnorm(mu2,tau2)}
#Prior
mu2~dnorm(310,0.1)
tau2~dgamma(0.0001,0.0001)
var2<-1/tau2
sd2<-1/sqrt(tau2)}

#M2_2.txt
model{
#Likelihood
for (i in 1:n2){
y2[i]~dnorm(mu2,tau2)}
#Prior
mu2~dnorm(310,2)
tau2~dgamma(0.0001,0.0001)
var2<-1/tau2
sd2<-1/sqrt(tau2)}

#M2_plana.txt
model{
#Likelihood
for (i in 1:n2){
y2[i]~dnorm(mu2,tau2)}
#Prior
mu2~dflat()
tau2~dgamma(0.0001,0.0001)
var2<-1/tau2
sd2<-1/sqrt(tau2)
```

##Estimación de una proporción

```
#Funciones necesarias para obtener a y b

#p media y tamaño muestral
ab_pn<-function(p,n){
  a<-p*(n-1)
  b<-n-a-1
  return(c(a,b))}

#intervalo de confianza al 95%
ab_pp<-function(p1,p2){
  p<-(p1+p2)/2
  n<-4*p*(1-p)/(p-p1)**2
  return(ab_pn(p,n))}

#p media y varianza
ayb_pv<-function(p,v){
  a<-p**2*(1-p)/v-p
  b<-p*(1-p)**2/v+p-1
  return(c(a,b))}
```

```
#Función para obtener la distribución posterior
beta_posterior<-function(a,b,y,n){
  a2<-a+y
  b2<-b+n-y
  p2<-a2/(a2+b2)
  return(c(a2,b2,p2))}

#Primer estrato

#a) p=0.3 n=50
ab_pn(0.3,50)#17.7, 41.3

#b) p=0.3 (0.2,0.4)
ab_pp(0.2,0.4)#24.9 58.1 (tamaño equivalente=83)

#c) Sin conocimiento
#a=1, b=1

#Simulación de datos datos, p=0,25
set.seed(1)
a<-runif(50,0,1)
x1<-rep(0,50)
for(i in 1:50){
  if(a[i]<=0.25){x1[i]<-1}}

x1<-factor(x1)
summary(x1)
#0 40
#1 10

#Despejamos a y b de la verosimilitud, p=0.2, n=50.
ab_pn(0.2,50) #9.8 39.2

#a) Previa 1

#Posterior
pos1<-beta_posterior(14.7,34.3,10,50)
pos1 #24.7000000 74.3000000  0.2494949

#gráfico 1

plot(z,dbeta(z,14.7,  34.3),ylab="",xlab="",cex.axis=2,col="blue",type="l",lwd  =
5,ylim=c(0,10))
title(main=expression("Conocimiento                 previo:                 p=0.3
n=50"),cex.main=2.1,xlab=expression(p[1.1]),cex.lab=3)
lines(z,dbeta(z,9.8, 39.2),ylab="",col="green",type="l",lwd = 5,ylim=c(0,8))
lines(z,dbeta(z,pos1[1], pos1[2]), main="" ,col="red",type="l",lwd = 5)
legend(0.45,8,nombres,pch=c(15,15,15),col=colores,cex=2)

#b) Previa 2

#Posterior
pos2<-beta_posterior(24.9,58.1,10,50)
pos2 #34.900000 98.100000  0.262406

#gráfico 2
```

```
plot(z,dbeta(z,24.9,  58.1),ylab="",xlab="",cex.axis=2,col="blue",type="l",lwd  =
5,ylim=c(0,11))
title(main=expression(paste("Conocimiento               previo:               ",
IC^{95~'%'},"=(0.2,0.4)")),cex.main=2.1,xlab=expression(p[1.2]),cex.lab=3)
lines(z,dbeta(z,9.8, 39.2),ylab="",col="green",type="l",lwd = 5,ylim=c(0,8))
lines(z,dbeta(z,pos2[1], pos2[2]), main="" ,col="red",type="l",lwd = 5)
legend(0.45,8,nombres,pch=c(15,15,15),col=colores,cex=2)

#c) Previa 3

#Posterior
pos3<-beta_posterior(1,1,10,50)
pos3 #11.0000000 41.0000000  0.2115385

#gráfico 3

plot(z,dbeta(z,1,     1),ylab="",xlab="",cex.axis=2,col="blue",type="l",lwd     =
5,ylim=c(0,11))
title(main=expression(paste("Sin                                          conocimiento
previo")),cex.main=2.1,xlab=expression(p[1.3]),cex.lab=3)
lines(z,dbeta(z,9.8, 39.2),ylab="",col="green",type="l",lwd = 5,ylim=c(0,8))
lines(z,dbeta(z,pos3[1], pos3[2]), main="" ,col="red",type="l",lwd = 5)
legend(0.45,8,nombres,pch=c(15,15,15),col=colores,cex=2)

#Segundo estrato

#a) p=0.2 n=35
ab_pn(0.2,35)#6.8 27.2

#b) p=0.2 (0.1,0.3)
ab_pp(0.1,0.3)#12.6 50.4 (tamaño equivalente=63)

#c) Sin conocimiento previo
#a=1, b=1

#Simulamos datos, p=0.1
set.seed(1)
a<-runif(30,0,1)
x2<-rep(0,30)
for(i in 1:30){
  if(a[i]<=0.15){x2[i]<-1}}

x2<-factor(x2)
summary(x2)
#0 27
#1 3

#Despejamos a y b, p=0.1 n=30
ab_pn(0.1,30) # 2.9 26.1

#a) Previa 1

#Posterior
pos2.1<-beta_posterior(6.8, 27.2,3,30)
pos2.1 #9.800000 54.200000  0.153125
```

```
#gráfico 1
plot(z,dbeta(z,6.8,   27.2),ylab="",xlab="",cex.axis=2,col="blue",type="l",lwd   =
5,ylim=c(0,11))
title(main=expression("Conocimiento                 previo:                 p=0.2
n=35"),cex.main=2.1,xlab=expression(p[2.1]),cex.lab=3)
lines(z,dbeta(z,2.9, 26.1),ylab="",col="green",type="l",lwd = 5,ylim=c(0,8))
lines(z,dbeta(z,pos2.1[1], pos2.1[2]), main="" ,col="red",type="l",lwd = 5)
legend(0.45,8,nombres,pch=c(15,15,15),col=colores,cex=2)

#b) Previa 2

#Posterior
pos2.2<-beta_posterior(12.6, 50.4,3,30)
pos2.2 #  15.6000000 77.4000000  0.1677419

#gráfico 2
plot(z,dbeta(z,12.6,  50.4),ylab="",xlab="",cex.axis=2,col="blue",type="l",lwd  =
5,ylim=c(0,11))
title(main=expression(paste("Conocimiento               previo:               ",
IC^{95~'%'},"=(0.1,0.3)")),cex.main=2.1,xlab=expression(p[2.2]),cex.lab=3)
lines(z,dbeta(z,2.9, 26.1),ylab="",col="green",type="l",lwd = 5,ylim=c(0,8))
lines(z,dbeta(z,pos2.2[1], pos2.2[2]), main="" ,col="red",type="l",lwd = 5)
legend(0.45,8,nombres,pch=c(15,15,15),col=colores,cex=2)

#c) Previa 3
#Posterior
pos2.3<-beta_posterior(1, 1,3,30)
pos2.3 # 4.000 28.000  0.125

#gráfico 3
plot(z,dbeta(z,1,     1),ylab="",xlab="",cex.axis=2,col="blue",type="l",lwd     =
5,ylim=c(0,11))
title(main=expression(paste("Sin conocimiento
previo")),cex.main=2.1,xlab=expression(p[2.3]),cex.lab=3)
lines(z,dbeta(z,2.9, 26.1),ylab="",col="green",type="l",lwd = 5,ylim=c(0,8))
lines(z,dbeta(z,pos2.3[1], pos2.3[2]), main="" ,col="red",type="l",lwd = 5)
legend(0.45,8,nombres,pch=c(15,15,15),col=colores,cex=2)
```

II. Eranskina

```
#Anexo II

#Incluimos la previa y el tamaño muestral equivalente

datos15$p_previa14<-NA

for(i in 1:dim(datos15)[1]){
  for(j in 1:dim(datos14)[1]){
    if(datos15$ESTR_DIRECTORIO[i]==datos14$ESTR_DIRECTORIO[j]){
      datos15$p_previa14[i]<-datos14$p_f[j]
      datos15$n_previa14[i]<-datos14$n_f[j]}
  }}

for(i in 1:dim(datos15)[1]){
  if(is.na(datos15$n_previa14[i])){datos15$n_previa14[i]=3}
  if(is.na(datos15$p_previa14[i])){datos15$p_previa14[i]=0.5}
}

#Calculamos parámetros a y b
a<-rep(0,h_15)
b<-rep(0,h_15)

for(i in 1:h){
  a[i]<- datos15$p_previa14[i]*(datos15$n_previa14[i]-1)
  b[i]<-datos15$n_previa14[i]-a[i]-1
  if (a[i]<=0) a[i]<-0.01
  if (b[i]<=0) b[i]<-0.01}

y<-datos15$y
n<-datos15$n
N<-datos15$N

data.previa<-list(y=y,n=n,N=N,h=h_15,a=a,b=b)

par.p <- c("p_total","pi")

inits<-function(){
  list(pi=runif(h,0,1))}

#Cargamos librería
library(R2WinBUGS)

set.seed(1)
bmbugs_previa <- bugs(data.previa,inits, par.p, "previa.txt",
                      n.chains = 1, n.iter = 10000, n.burnin=1000,
                      bugs.directory ="D:/DATOS/NGUTIEAR/Desktop/WinBUGS14/")

print(bmbugs_previa,digits=4)


#Simulaciones de cada estrato
pi_post_previa_h<-matrix(0,h,1000)
for(i in 1:h){
  pi_post_previa_h[i,]<-bmbugs_previa$sims.array[,1 ,paste("pi[",i,"]",sep="")]}
```

```
#Gipuzkoa

p_total_previa_b<- bmbugs_previa$sims.array[,1 ,"p_total"]

p_tot_previa_b<-mean(p_total_previa_b)
p_tot_previa_b
#  0.2659584

l_previa_b<-quantile(p_total_previa_b, c(0.025))
l_previa_b
#0.2464925

u_previa_b<-quantile(p_total_previa_b, c(0.975))
u_previa_b
#0.2845025

#Por sectores A38

datos15$ACTA38<-factor(datos15$ACTA38)

a38<-datos15$ACTA38

dominios<-list("1","2","3","4","5","6","7","8","9","10","11","12",
               "13","14","15","16","17","18","19","20","21","22","23","24",
               "25","26","27","28","29","30","31","32","33","34","35","36",
               "37","38")

d<-length(dominios)

p_dom1_previa<-matrix(0,d,1000)
p_dom1_med_previa<-rep(0,d)
for(i in 1:d){
  t<-rep(0,1000)
  tot<-0
  for(j in 1:h){
    if (dominios[i]==a38[j]){
      t<-t+pi_post_previa_h[j,]*N[j]
      tot<-tot+N[j]
    }
    p_dom1_previa[i,]<-t/tot
    p_dom1_med_previa[i]<-mean(t/tot) }}


names(p_dom1_med_previa)<-
list("1","2","3","4","5","6","7","8","9","10","11","12",

"13","14","15","16","17","18","19","20","21","22","23","24",

"25","26","27","28","29","30","31","32","33","34","35","36",
                               "37","38")

p_d1_previa<-matrix(0,length(na.omit(p_dom1_med_previa)),2)
for(i in 1:length(na.omit(p_dom1_med_previa))){
  p_d1_previa[i,1]<-as.numeric(names(na.omit(p_dom1_med_previa)))[i]
  p_d1_previa[i,2]<-na.omit(p_dom1_med_previa)[i]}

p_d1_previa
```

```
#A38="02"
p_A38_2_previa_b<-p_dom1_med_previa[2]
p_A38_2_previa_b
#0.00169441

l_A38_2_previa_b<-quantile(p_dom1_previa[2,], c(0.025))
u_A38_2_previa_b<-quantile(p_dom1_previa[2,], c(0.975))
quantile(p_dom1_previa[2,], c(0.025,0.975))
#0.00000000 0.01596836

#A38="03"
p_A38_3_previa_b<-p_dom1_med_previa[3]
p_A38_3_previa_b
#0.3280667

l_A38_3_previa_b<-quantile(p_dom1_previa[3,], c(0.025))
u_A38_3_previa_b<-quantile(p_dom1_previa[3,], c(0.975))
quantile(p_dom1_previa[3,], c(0.025,0.975))
#0.1990041 0.4694112

#A38="04"
p_A38_4_previa_b<-p_dom1_med_previa[4]
p_A38_4_previa_b
#0.240075

l_A38_4_previa_b<-quantile(p_dom1_previa[4,], c(0.025))
u_A38_4_previa_b<-quantile(p_dom1_previa[4,], c(0.975))
quantile(p_dom1_previa[4,], c(0.025,0.975))
#0.1281434 0.3654042

#A38="05"
p_A38_5_previa_b<-p_dom1_med_previa[5]
p_A38_5_previa_b
#0.3814882

l_A38_5_previa_b<-quantile(p_dom1_previa[5,], c(0.025))
u_A38_5_previa_b<-quantile(p_dom1_previa[5,], c(0.975))
quantile(p_dom1_previa[5,], c(0.025,0.975))
#0.2504197 0.5134477

#A38="07"
p_A38_7_previa_b<-p_dom1_med_previa[7]
p_A38_7_previa_b
#0.3585053

l_A38_7_previa_b<-quantile(p_dom1_previa[7,], c(0.025))
u_A38_7_previa_b<-quantile(p_dom1_previa[7,], c(0.975))
quantile(p_dom1_previa[7,], c(0.025,0.975))
#0.2276541 0.5180853

#A38="09"
p_A38_9_previa_b<-p_dom1_med_previa[9]
p_A38_9_previa_b
#0.2173366

l_A38_9_previa_b<-quantile(p_dom1_previa[9,], c(0.025))
```

```
u_A38_9_previa_b<-quantile(p_dom1_previa[9,], c(0.975))
quantile(p_dom1_previa[9,], c(0.025,0.975))
#0.1321552 0.3167207

#A38="10"
p_A38_10_previa_b<-p_dom1_med_previa[10]
p_A38_10_previa_b
#0.3001016

l_A38_10_previa_b<-quantile(p_dom1_previa[10,], c(0.025))
u_A38_10_previa_b<-quantile(p_dom1_previa[10,], c(0.975))
quantile(p_dom1_previa[10,], c(0.025,0.975))
#0.2131268 0.3986713

#A38="11"
p_A38_11_previa_b<-p_dom1_med_previa[11]
p_A38_11_previa_b
#0.4068358

l_A38_11_previa_b<-quantile(p_dom1_previa[11,], c(0.025))
u_A38_11_previa_b<-quantile(p_dom1_previa[11,], c(0.975))
quantile(p_dom1_previa[11,], c(0.025,0.975))
#0.3138329 0.5084525

#A38="13"
p_A38_13_previa_b<-p_dom1_med_previa[13]
p_A38_13_previa_b
#0.3319086

l_A38_13_previa_b<-quantile(p_dom1_previa[13,], c(0.025))
u_A38_13_previa_b<-quantile(p_dom1_previa[13,], c(0.975))
quantile(p_dom1_previa[13,], c(0.025,0.975))
#0.2410149 0.43596551

#A38="14"
p_A38_14_previa_b<-p_dom1_med_previa[14]
p_A38_14_previa_b
#0.3459527

l_A38_14_previa_b<-quantile(p_dom1_previa[14,], c(0.025))
u_A38_14_previa_b<-quantile(p_dom1_previa[14,], c(0.975))
quantile(p_dom1_previa[14,], c(0.025,0.975))
#0.2207378 0.4830781

#A38="15"
p_A38_15_previa_b<-p_dom1_med_previa[15]
p_A38_15_previa_b
#0.3365971

l_A38_15_previa_b<-quantile(p_dom1_previa[15,], c(0.025))
u_A38_15_previa_b<-quantile(p_dom1_previa[15,], c(0.975))
quantile(p_dom1_previa[15,], c(0.025,0.975))
#0.2222601 0.4636895

#A38="16"
p_A38_16_previa_b<-p_dom1_med_previa[16]
p_A38_16_previa_b
#0.1259951
```

```
l_A38_16_previa_b<-quantile(p_dom1_previa[16,], c(0.025))
u_A38_16_previa_b<-quantile(p_dom1_previa[16,], c(0.975))
quantile(p_dom1_previa[16,], c(0.025,0.975))
#0.0176104 0.3052162

#A38="17"
p_A38_17_previa_b<-p_dom1_med_previa[17]
p_A38_17_previa_b
#0.1172628

l_A38_17_previa_b<-quantile(p_dom1_previa[17,], c(0.025))
u_A38_17_previa_b<-quantile(p_dom1_previa[17,], c(0.975))
quantile(p_dom1_previa[17,], c(0.025,0.975))
#0.04320774 0.22660154

#A38="18"
p_A38_18_previa_b<-p_dom1_med_previa[18]
p_A38_18_previa_b
#0.1628718

l_A38_18_previa_b<-quantile(p_dom1_previa[18,], c(0.025))
u_A38_18_previa_b<-quantile(p_dom1_previa[18,], c(0.975))
quantile(p_dom1_previa[18,], c(0.025,0.975))
#0.1014900 0.2340487

#A38="19"
p_A38_19_previa_b<-p_dom1_med_previa[19]
p_A38_19_previa_b
#0.3163508

l_A38_19_previa_b<-quantile(p_dom1_previa[19,], c(0.025))
u_A38_19_previa_b<-quantile(p_dom1_previa[19,], c(0.975))
quantile(p_dom1_previa[19,], c(0.025,0.975))
#0.2701332 0.3631196

#A38="20"
p_A38_20_previa_b<-p_dom1_med_previa[20]
p_A38_20_previa_b
#0.1378989

l_A38_20_previa_b<-quantile(p_dom1_previa[20,], c(0.025))
u_A38_20_previa_b<-quantile(p_dom1_previa[20,], c(0.975))
quantile(p_dom1_previa[20,], c(0.025,0.975))
#0.08092351 0.21271029

#A38="21"
p_A38_21_previa_b<-p_dom1_med_previa[21]
p_A38_21_previa_b
#0.1499432

l_A38_21_previa_b<-quantile(p_dom1_previa[21,], c(0.025))
u_A38_21_previa_b<-quantile(p_dom1_previa[21,], c(0.975))
quantile(p_dom1_previa[21,], c(0.025,0.975))
#0.1095624 0.2013698

#A38="22"
p_A38_22_previa_b<-p_dom1_med_previa[22]
```

```
p_A38_22_previa_b
#0.4769936

l_A38_22_previa_b<-quantile(p_dom1_previa[22,], c(0.025))
u_A38_22_previa_b<-quantile(p_dom1_previa[22,], c(0.975))
quantile(p_dom1_previa[22,], c(0.025,0.975))
#0.3978893 0.5464147

#A38="25"
p_A38_25_previa_b<-p_dom1_med_previa[25]
p_A38_25_previa_b
#0.1810702

l_A38_25_previa_b<-quantile(p_dom1_previa[25,], c(0.025))
u_A38_25_previa_b<-quantile(p_dom1_previa[25,], c(0.975))
quantile(p_dom1_previa[25,], c(0.025,0.975))
#0.1174039 0.2537635

#A38="26"
p_A38_26_previa_b<-p_dom1_med_previa[26]
p_A38_26_previa_b
#0.2026248

l_A38_26_previa_b<-quantile(p_dom1_previa[26,], c(0.025))
u_A38_26_previa_b<-quantile(p_dom1_previa[26,], c(0.975))
quantile(p_dom1_previa[26,], c(0.025,0.975))
#0.1174303 0.3137936

#A38="27"
p_A38_27_previa_b<-p_dom1_med_previa[27]
p_A38_27_previa_b
#0.3873074

l_A38_27_previa_b<-quantile(p_dom1_previa[27,], c(0.025))
u_A38_27_previa_b<-quantile(p_dom1_previa[27,], c(0.975))
quantile(p_dom1_previa[27,], c(0.025,0.975))
#0.3262329 0.4522949

#A38="29"
p_A38_29_previa_b<-p_dom1_med_previa[29]
p_A38_29_previa_b
#0.222496

l_A38_29_previa_b<-quantile(p_dom1_previa[29,], c(0.025))
u_A38_29_previa_b<-quantile(p_dom1_previa[29,], c(0.975))
quantile(p_dom1_previa[29,], c(0.025,0.975))
#0.1349709 0.3286571

#A38="30"
p_A38_30_previa_b<-p_dom1_med_previa[30]
p_A38_30_previa_b
#0.3744737

l_A38_30_previa_b<-quantile(p_dom1_previa[30,], c(0.025))
u_A38_30_previa_b<-quantile(p_dom1_previa[30,], c(0.975))
quantile(p_dom1_previa[30,], c(0.025,0.975))
#0.2334885 0.5236972
```

```
#A38="31"
p_A38_31_previa_b<-p_dom1_med_previa[31]
p_A38_31_previa_b
#0.1787776

l_A38_31_previa_b<-quantile(p_dom1_previa[31,], c(0.025))
u_A38_31_previa_b<-quantile(p_dom1_previa[31,], c(0.975))
quantile(p_dom1_previa[31,], c(0.025,0.975))
#0.1006409 0.2836296

#A38="32"
p_A38_32_previa_b<-p_dom1_med_previa[32]
p_A38_32_previa_b
#0.3581569

l_A38_32_previa_b<-quantile(p_dom1_previa[32,], c(0.025))
u_A38_32_previa_b<-quantile(p_dom1_previa[32,], c(0.975))
quantile(p_dom1_previa[32,], c(0.025,0.975))
#0.2809560 0.4493003

#A38="33"
p_A38_33_previa_b<-p_dom1_med_previa[33]
p_A38_33_previa_b
#0.342817

l_A38_33_previa_b<-quantile(p_dom1_previa[33,], c(0.025))
u_A38_33_previa_b<-quantile(p_dom1_previa[33,], c(0.975))
quantile(p_dom1_previa[33,], c(0.025,0.975))
#0.2383233 0.4555771

#A38="34"
p_A38_34_previa_b<-p_dom1_med_previa[34]
p_A38_34_previa_b
#0.1927656

l_A38_34_previa_b<-quantile(p_dom1_previa[34,], c(0.025))
u_A38_34_previa_b<-quantile(p_dom1_previa[34,], c(0.975))
quantile(p_dom1_previa[34,], c(0.025,0.975))
#0.1515279 0.2396318

#previa.txt
model{

#likelihood
for(i in 1:h){
y[i]~dbin(pi[i],n[i])}

#priors
for(i in 1:h){
pi[i]~dbeta(a[i],b[i])}

#p_total
for(i in 1:h){
m[i]<-pi[i]*N[i]}

p_total<-sum(m[1:h])/sum(N[1:h])}
```

Bibliografia

R Core Team. (2015). R: A language and environment for statistical. R Foundation for Statistical Computing, Vienna, Austria.: URL: http://www.R-project.org/.

Bolstad, J. C. (2016). Bolstad: Bolstad functions. R package version 0.2-33. .

Bolstad, W. M. (2007). Introduction to Bayesian Statistics. 2nd edtition. WILEY.

Congdon, P. (2006). Bayesian Statistical Modelling. WILEY.

EUSTAT. (2011). Proyecto Técnico de la Operación Encuesta sobre Sociedad de la Información en las empresas (ESIE).

EUSTAT. (2012). Estimación de áreas pequeñas en la encuesta sobre la sociedad de la información en las empresas de C.A. de Euskadi.

Gelman A., Carlin J. B., Sten H. S., Dunson D. B. , Vehtar A.i and Rubin D. B. . (2013). Bayesian Data Analysis. Third Edition. CRC Press.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis , 1, Number 3, pp. 515–533.

Hoff, P. D. (2009). A First Course in Bayesian Statistical Methods. Springer.

Kruschke, J. K. (2015). Doing Bayesian Data Analysis. A tutorial with R, JAGS and Stan. ACADEMIC PRESS.

Little, R. (2016). Bayesian inference for sample surveys. University of Michigan.

Lunn D., Jackson C., Best N., Thomas A. & Spiegelhalter D. (2012). The BUGS Book: A Practical Introduction to Bayesian Analysis. CRC Press.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325--337.

Ntzoufras, I. (2008). Bayesian Modeling Using WinBUGS. WILEY.

Statistics Canada. (2014). Survey Methodology. A Journal Published by Statistics Canada., Volume 40, Number 1.

Sturtz, S. L. (2005). R2WinBUGS: A Package for. Journal of Statistical Software, 12(3), 1-16.

Organismo Autónomo del

www.eustat.eus