

Inferentziarako sarrera

Josemari Sarasola

Estatistika enpresara aplikatua

Gizapedia

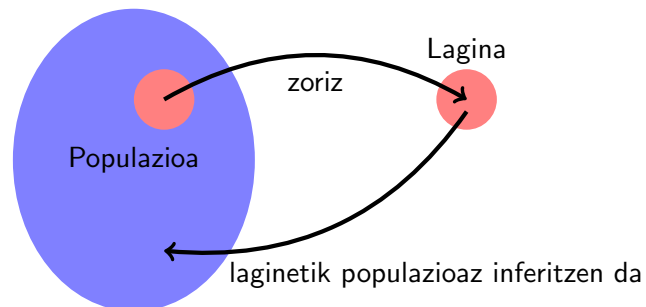


Inferentzia estatistikoa

Estatistikan, populazioak (adibidez, 18 urteko gazteak herrialde batean) ikertu nahi ditugu, zehatzago populazioari buruzko ezaugarri bat (altuera, adibidez). Ezaugarri hori aldakorra izaten da, zorizkotzat har dezakegu, eta beraz probabilitate banaketa bat dagokio, parametro zenbaitekin. Probabilitate banaketa hori populazioaren adierazpen sinplea da, eta beraz populazioaren eredia dela esan dezakegu. Hurrengo azalpenetan, *populazioa*, *probabilitate banaketa eta eredia* (ia) sinonimotzat har ditzakegu.

Inferentzia estatistikoa

Gehienetan, ezin ditugu aztertu populazio osoko elementuak (garestiegia delako edo populazioko elementu guztiak zerrendatu ezin direlako), eta orduan lagin bat hartzen da populazioari buruzko informazioa izateko (populazioari buruz inferitu ahal izateko). Laginak *zoriz hartutakoak* izan behar dira, populazioaren adierazgarri izango badira.



Inferentzia estatistikoa

Aurreko ikasgaietan, eredu estatistikoek (binomial, Poisson, esponentziala normala) aplikazio interesgarriak modu sinplean ebazteko balio dutela ikasi dugu. Horretarako, eredu horietako parametroak kuantifikatuta izatea ezinbestekoa da.

Baino bi galdera hauek planteatzen dira orduan: nola zehazten dugu probabilitate banaketa bat populazio baterako? Eta nola zehazten ditugu bere parametroak?

- 1. galderaren erantzuna: hasiera batean eredia edo probabilitate banaketa (parametroak kuantifikatu gabe) suposatu egiten da (gutxi gorabehera, histogramari erreparatuz adibidez: kanpai itxurakoa bada, eredu normala esleitzen diogu populazioari).
- 2. galderaren erantzuna: Parametroak lagineko datuetatik kuantifikatu behar dira, estimatzaileak erabiliz, lagin datuetatik kalkulatu diren formulak (adibidez, Poisson populazio baten lambda parametroa kuantifikatzeko, datuen batezbestekoa kalkulatu).

Problema nagusia

Problema nagusia **inferentzia estatistikoa**n hau da: populazio bateko banaketa-parametroak **inferitu** populazio osotik jasotako **laginak** baliatuz.

Laginak populazio batetik erautzen diren datu-azpimultzoak dira. Ikergai den multzo osoa populazioa da, baina hura datuz datu osorik jasotzea ezinezkoa denez, hortik datu batzuk soilik jasotzen dira, lagina osatzen dutenak. Lagina populazioaren adierazgarria izateko, datuak zoriz jaso behar dira, zehatzago *zorizko laginketa sinpleaz*, non populazioko elementu guztiek aukeratuak izateko probabilitate berdina duten, eta beraz zoriz eta independentziaz aukeratzeko diren.

Datuak bildurik, datuei dagokien eredia erabaki behar da. Bi erataraz egin daitezke:

- datuen histograma edo beste grafiko bat eratuz, nolako itxura duen ikusita. Adibidez, aski da datuen histogramak kanpai itxurakoa izatea banakuntza normala onartzeko;
- datuen izaerari berari erreparatuz: datuak bai/ez motakoak badira, eredu binomiala da oinarritzakoa, independentzia suposatuz.

Lagineko datuekin, estimatzailak (zenbatesleak edo estatistikoak ere deituak) kalkulatu dira, populazioko ezaugarriak, ereduko parametroak alegia, kuantifikatzen. Orohar, parametroari θ (theta) deitzen zaio, eta horren estimatzaile edo zenbatesle bati $\hat{\theta}$. Ohartu behar da parametroak (txanorik gabe) orokorrean ezezagunak izango direla, baita estimatu ondoren ere; estimatzailak horien zenbatespen edo estimazioak (txanoarekin) emango dituzte.

Adibidez, populazioko (ereduko) batezbestekoa (*population mean*) zenbatetsi edo estimatzeko, lagin batezbestekoa (*sample mean*) kalkulatu ohi da. Zenbatespen edo estimazio hori honela idazten da: $\hat{\mu} = \bar{x}$. Estimatzaila intuitibo horiei *estimatzaila natural* deitzen zaie.

Parametro baterako propietate *onak* dituzten estimatzailak aukeratu behar da, orohar haren balio ezezagunetik gertu ibiliko dena.

Inferentzia estatistikoa

4. fasea: Parametroen kuantifikazioa

Parametro baterako estimatzaileak aukeratuta, bi modutara kuantifikatu daitezke parametroak:

- zuzeneko estimazioaz, hau da, parametroaren balioztat estimatzaileak ematen duena hartuz; adibidez, $\hat{\mu} = \bar{x} = 4.5$.
- parametroaren balio zehatz bat hipotesi nulutzat hartuz eta estimatzailearen balioa ikusita erabakiz hipotesi hori bazter daitekeen ala ez; adibidez, $H_0 : \mu = 4$ onartu egiten dut \bar{x} estimatzailearen emaitza ikusita.

Inferentzia estatistikoa

Parametroen eta estimatzaileen arteko diferentziak

Parametroak	Estimatzaileak
Notazioa: θ Populazioari edo ereduari dagozkio Konstanteak dira Ezezagunak izaten dira θ bakarra da Adibidea: μ (populazioko batezbestekoa)	Notazioa: $\hat{\theta}$ (θ -ren estimatzailea) Laginari dagozkio Aldakorrak dira Kalkulatu egiten dira datuetatik $\hat{\theta}$ estimatzaile anitz daude eskura Adibidea: $\hat{\mu}_1 = \bar{x}$, $\hat{\mu}_2 = Me$

Ohikoak dira *estimatzaile naturalak*, populazioko parametroak estimatzeko laginean oinarritutako formula baliokideak alegia. Adibidez, populazioaren μ estimatzaile naturala lagineko datuen $\hat{\mu} = \bar{x}$ da.

Inferentzia estatistikoa

5. fasea eta azkena: Balidazioa

Parametroak kuantifikatuta, ordurarte egindako guztia zuzena den frogatu behar da, hau da, ereduaren balidazioa egin behar da.

Zehatzago:

- datuak benetan zoriz jaso diren;
- aurrez suposatu dugun eredu edo banaketa benetan egokia den jasotako datuetarako (doikuntzaren egokitasuna, ingelesez *goodness of fit*)
- datuak homogeneousak diren, hots, populazio bakar bati buruzkoak diren (adibidez, emakume eta gizonen datuak batera jarri, frogatu behar da bi sexuak benetan berdina diren ala ez, datuak lagin berean bildu ahal izateko).

Jarraian, balidazioaren hiru alderdi hauek nola frogatu ikasi behar dugu.

Balidazioa

Zorizkotasunerako froga: bolada-froga

Aldagai dikotomikoetarako eta kuantitatiboetarako (balio bakoitza medianatik gora edo behera dagoen jarrita) gara daiteke, datuak zoriz (eta beraz independentziaz) jaso diren erabakitzeko. Bolada bat balio edo zeinuko bereko datu segidetako bakoitza da. Datuak kuantitatiboak direnean, datu segidak medianatik gora eta behera dauden adierazten dute boladek. Froga datu-multzo osoan dagoen bolada-kopuruan oinarritzen da. Adibidez, XX0XX000XX akastun eta akasgabeen sekuentzian bolada kopurua 5 da.

Balidazioa

Zorizkotasunerako froga: bolada-froga

Hiru egoera posible:

- XXXXXOOOOO: 2 bolada (hots, gutxi) → zorizkotasun-eza edo dependentzia
- XOXOXOXOXO: 10 bolada (hots, asko) → zorizkotasun-eza edo dependentzia
- XXOOOXOXXO: 6 bolada (hots, ez asko ez gutxi) → zorizkotasuna, eta beraz independentzia

Beraz, [H_0 : independentzia] bolada kopurua oso handia edo oso txikia denean baztertzen da.

Balidazioa

Zorizkotasunerako froga: bolada-froga

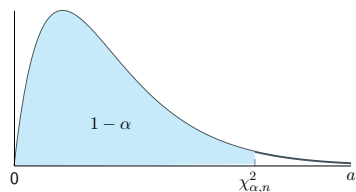
- Datuak jaso diren ordenan hartu behar dira beti.
- Froga alde bikoa da, H_0 -pean arraroa goitik nahiz behetik dagoenez.
- Erabakia hartzeko taulak erabiltzen dira, *balio kritikoak* ematen dituena
- Datu kopurua handia denean, boladak honela banatzen dira H_0 -pean:

$$R \sim N\left(\mu = \frac{2n_1n_2}{n_1 + n_2} + 1, \sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}\right)$$

Balidazioa

Doikuntzaren egokitasunerako froga: khi-karratu banaketa

Froga honetan, probabilitate banakuntza berezi bat baliatu behar dugu: χ_n^2 (khi-karratu) banakuntza, n parametro bakarra duena (*askatasun-mailak* izenekoa eta zenbaki naturala izan behar dena), eta balio positiboak soilik hartzen dituena. Honelakoa izaten da, eskubirantz alboratua:



Bere balioak taularatuta daude, $1 - \alpha$ azpiko probabilitate zehatzetarako. Adibidez,

- $\chi_{0.01,4}^2 = 13.3$
- $\chi_{0.25,2}^2 = 2.77$

Balidazioa

Doikuntzaren egokitasunerako froga: khi-karratu froga

- H_0 : eredia zuzena da, kuantifikatutako parametroekin.
- Aldagaiaren balio edo balio-tarte bakoitzeko maiztasun enpirikoak (O_i) eta teorikoak (E_i), azken horiek ereduko probabilitateetatik, kalkulatu dira.
- $X^2 = \frac{(O_i - E_i)^2}{E_i}$ estatistikoa kalkulatu.
- X^2 oso handia denean, maiztasun teorikoen eta enpirikoen arteko aldea handia da, eta beraz eredia zuzena ez dela esateko joera beharko genuke izan. Froga alde bakarrekoa da eta *arraroa goitik dago*, hortaz.
- Froga burutzeko, X^2 estatistikoaren emaitza balio kritikoarekin alderatu behar da:
 - $\chi_{\alpha, k-1}^2$ balioarekin, k izanik balio edo tarte desberdinen kopurua; edota,
 - parametroak estimatu direnean, $\chi_{\alpha, k-z-1}^2$ balioarekin, z izanik datuetatik *estimatu* diren parametroen kopurua.

Balidazioa

Homogeneotasunerako: Wilcoxon hein-froga

- Datu kuantitatiboetarako erabiltzen da, atributu dikotomiko baten arabera bereiz daitezkeenak.
- H_0 : atributuak ez du alderik eragiten \rightarrow homogeneotasuna
- Datu guztiak txikienetik handienara ordenatu.
- Heinak (mailak edo ordenak) jarri, atributuaren bi kategorien arabera bereizita.
- Atributuko kategoria bakoitzeko, W heinen batura kalkulatu. Bietatik txikiena hartu, W_{min} deituko duguna.

Balidazioa

Homogeneotasunerako: Wilcoxon hein-froga

- W_{min} oso txikia denean, bi datu azpimultzoak oso desberdinak direla esan nahi du. Froga alde bikoia da, minimoa bi taldeetako edozeini egoki dakiokelako, baina minimoa hartzen dugunez beti behetik begiratzen da.
- Balio kritikoak taulan bilatzen dira, kategoria bietako datu-kopuruaren arabera, lagin txikietarako ($n_1, n_2 \leq 20$).
- Lagin handietarako ($n_1, n_2 > 20$), honela banatzen da estatistikoa, n_1 1 kategoriako lagin-tamaina izanik:

$$W_1 \sim N\left(\mu = \frac{n_1(n_1 + n_2 + 1)}{2}, \sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}\right)$$

AMAIERA