Inferentziarako sarrera

Josemari Sarasola

Estatistika enpresara aplikatua

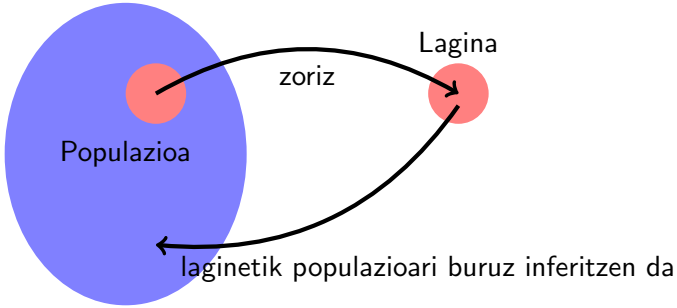
Gizapedia

Inferentzia estatistikoa

Estatistikan, populazioak (adibidez, 18 urteko gazteak herrialde batean) ikertu nahi ditugu, kolektibo gisa eta horren ezaugarri jakin bati buruz (18 urteko gazteen altuera) edo aldagai mugababe gisa (eguneko ekoizpena lantegi batean). Era batera zein bestera, ezaugarri eta aldagai horiek aldakorrak izaten dira, zorizkotzat har ditzakegu, eta beraz probabilitate banaketa bat dagokie. Probabilitate banaketa populazioaren eredu edo adierazpen modelizatua dela esan dezakegu, beraz. Hurrengo azalpenetan, *populazioa, probabilitate banaketa eta eredua* (ia) sinonimotzat hartuko ditugu.

Inferentzia estatistikoa

Gehienetan, ezin ditugu aztertu populazio osoko elementuak (garestiegia delako edo, aldagaia mugagabea denean -populazioa infinitua denean- 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

- Inferentziaren helburua (*inferitu* aditzetik, "konklusioak atera") lagin bateko datuetatik populazioko edo dagokien probabilitate banaketako parametroak kuantifikatu edo haiei buruzko konklusioak ateratzea da.
- Nola kuantifikatzen dira parametroak? Parametroak edo populazioaren ezaugarri konkretuak lagineko datuetatik kuantifikatzen dira, estimatzaileak edo estatistikoak, datuetan oinarritutako formulak alegia, erabiliz (adibidez, Poisson populazio baten λ parametroa kuantifikatzeko, datuen batezbestekoa kalkulatuz).
- Parametro horiek kuantifikatuta, problema praktiko interesgarriak ebatz ditzakegu, probabilitate banaketei buruzko aurreko ikasgaietan ikusi dugunez.

Inferentzia estatistikoaren faseak

(1) Lagina jasotzea.
(2) Eredua aukeratzea.
(3) Estimatzaileak edo estatistikoak finkatzea.
(4) Parametroak kuantifikatzea.
(5) Balidazioa edo baliozkotzea

Inferentzia estatistikoaren faseak

1. fasea: Lagina jasotzea

Laginak populazio batetik erauzten 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 aukeratzen diren.

Inferentzia estatistikoaren faseak

2. fasea: Eredua aukeratzea

Datuak bildurik, datuei dagokien eredua erabaki behar da. Bi eratara egin daiteke:

- 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 aukeratzea da ohikoena, independentzia suposatuz.

Inferentzia estatistikoaren faseak

3. fasea: Estimatzaileak finkatzea

• Lagineko datuekin, estimatzaileak edo zenbatesleak (batzuetan estatistikoak ere deituak) kalkulatzen dira, populazioko ezaugarriak, ereduko parametroak alegia, kuantifikatzearren.
• Orohar, parametro bati 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; estimatzaileek horien zenbatespen edo estimazioak (txanoarekin) baino ez dira.
• Adibidez, populazioko (ereduko) batezbestekoa (*population mean*) zenbatetsi edo estimatzeko, lagin batezbestekoa (*sample mean*) kalkulatu ohi da. Zenbatespen edo estimazio hori honela idazten da: $\hat{\mu} = \overline{x}$. Estimatzaile intuitibo horiei *estimatzaile natural* deitzen zaie.
• Parametro baterako propietate *onak* dituzten estimatzailea aukeratu behar da, orohar haren balio ezezagunetik gertu ibiliko dena.

Inferentzia estatistikoaren faseak

4. fasea: Parametroen kuantifikazioa

Parametro baterako estimatzaileak aukeratuta, bi modutara kuantifikatu daitezke parametroak:

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

Inferentzia estatistikoa

Parametroen eta estimatzaileen arteko diferentziak

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

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

Inferentzia estatistikoaren faseak

5. fasea eta azkena: Balidazioa edo baliozkotzea

Parametroak kuantifikatuta, ordurarte egindako guztia zuzena den frogatu behar da, hau da, ereduaren balidazioa egin behar da. Zehatzago, hauek probatu edo egiaztatu behar dira:

(1) aurrez suposatu dugun eredua edo banaketa benetan egokia den jasotako datuetarako (doikuntzaren egokitasuna, ingelesez *goodness of fit* aztertu behar da, doikuntza-proben bitartez);

(2) datuak zoriz eta independentziaz jaso diren;

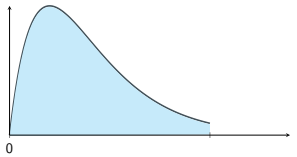
(3) datuak homogeneoak diren, hots, populazio bakar bati buruzkoak diren eta beraz datu guztiak batera jar daitezkeen (adibidez, emakume eta gizonen datuak batera jarririk, probatu behar da bi sexuak benetan berdinak diren ala ez, datuak lagin berean bildu ahal izateko).

Oharra: zorizkotasuna eta homogeneotasuna parametroen estimazioaren aurretik probatu daitezke, baino doikuntza-probak inferentzia eta gero garatu behar dira ezinbestean.

Balidazioa

Doikuntza probak: khi-karratu banaketa

Proba honetan, probabilitate banaketa berezi bat baliatu behar dugu: χ^2_n (khi-karratu) banaketa, n parametro bakarra duena (*askatasun-mailak edo graduak* 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,

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

Balidazioa

Doikuntza-probak: khi-karratu proba

- H_0: eredua zuzena da, kuantifikatutako parametroekin.
- Aldagaiaren balio edo balio-tarte bakoitzeko maiztasun enpirikoak edo behatuak (O_i, observed) eta teorikoak edo itxarondakoak (E_i, expected), azken horiek ereduko probabilitateetatik, kalkulatzen 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 eredua zuzena ez dela esateko joera beharko genuke izan. Proga alde bakarrekoa da eta *arraroa goitik dago*, hortaz.
- Proba burutzeko, X^2 estatistikoaren emaitza balio kritikoarekin alderatu behar da:
  - $\chi^2_{\alpha,k-1}$ balioarekin, k izanik balio edo tarte desberdinen kopurua; edota,
  - parametroak estimatu direnean, $\chi^2_{\alpha,k-z-1}$ balioarekin, z izanik datuetatik *estimatu* diren parametroen kopurua.

Balidazioa

Doikuntza-probak: khi-karratu proba

Adibidea

Txanpon bat 200 aldiz bota eta 86 aurpegiko eta 114 gurutzeko suertatu dira. %10eko adierazgarritasun-mailaz, txanpona orekatua baiezta al daiteke?
Eredua: p(o)=p(x)=0.5

Klaseak	Enpirikoak (O)	Prob.	Teorikoak (E)	$\frac{(O-E)^2}{E}$
o	86	0.5	$0.5 \times 200 = 100$	1.96
x	114	0.5	$0.5 \times 200 = 100$	1.96
	200		200	$\mathbf{X^2 = 3.92}$

2-1 askatasun-graduko khi-karratu banaketan, %10eko probabilitatea gainetik uzten duen balioa 2.71 da. Beraz, estatistikaren balioa (3.92, enpiriko-teoriko distantzia) esanguratsua da eta txanpon orekatuaren eredua baztertu egiten da.

Balidazioa

Zorizkotasuna: bolada-proba

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. Proba datu-multzo osoan dagoen bolada-kopuruan oinarritzen da. Adibidez, XX0XX000XX akastun eta akasgabeen segidan bolada kopurua 5 da.

Balidazioa

Zorizkotasuna: bolada-proba

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

Zorizkotasuna: bolada-proba

- Datuak jaso diren ordenan hartu behar dira beti.
- Proba 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

Homogeneotasuna: Wilcoxon hein-proba

- Datu kuantitatiboetarako erabiltzen da, atributu dikotomiko baten arabera bereiz daitezkeenak.
- H_0: atributuak ez du alderik eragiten $\rightarrow$ homogeneotasuna
- Datu guztiak txikienetik handienera 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

Homogeneotasuna: Wilcoxon hein-proba

- W_{min} oso txikia denean, bi datu azpimultzoak oso desberdinak direla esan nahi du. Proba alde bikoa da, minimoa bi taldeetako edozeini egoki dakiokeelako, 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:

$$W_1 \sim N\left(\mu = \frac{n_1(n_1+n_2+1)}{2},\ \sigma = \sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}\right)$$

AMAIERA