Estatistika Inferentziala
R software librea erabiliz

5. Hipotesi-kontraste ez-parametrikoak

**Artxibo honetako irudi guztiak irakasle taldeak prestatutako irudi propioak dira.*

Eneko Arrospide, Gorka Bidegain, Xabier Erdocia, Aitziber Unzueta

Universidad del País Vasco
Euskal Herriko Unibertsitatea

AURKIBIDEA

5.1. Sarrera
5.2. Independentzia probak
5.3. Normaltasun probak
    5.3.1. Kolmogórov-Smirnov
    5.3.2. Shapiro-Wilk

5.1. Sarrera

- Kurtso honetako aurreko ataletan populazioaren parametroei buruzko estimazioak eta hipotesi-kontrasteak egin dira.
- Horiek egin ahal izateko hainbat suposizio egin dira, ala nola, populazioaren banaketa normala zela edo bi aldagaien independentzia suposatzea.
- Gai honetan, suposizio horiek betetzen diren edo ez konprobatzeko metodoak erakutsiko dira.
- Metodo hauek hipotesi-kontrasteetan oinarritzen dira, baina ez dira populazioko parametroei buruzkoak baizik eta populazioaren banaketa zein den eta bi aldagaien arteko menpekotasuna edo independentzia dagoen onartzea ahalbidetuko digute.

5.2. Independentzia probak

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

- Bi ezaugarri edo faktore (aldagai) populazio batean duten independentzia edo mendekotasuna aztertzea alor askotan ohiko prozedura bat da.

- Ezaugarri horien arteko independentzia aztertzeko hipotesi kontraste ez-parametriko bat burutzen da non kontrasterako estatistikoak χ^2 banaketa bat jarraitzen duela hurbildu daitekeen. Beraz, **independentzia proba** hau **χ^2 -ren aplikazio bat**en moduan definitu daiteke.

Bi aldagaien independentziarako χ^2-ren proba

- Demagun n elementuz osatutako lagin bat non elementu bakoitza bi ezaugarriekiko (X eta Y) klasifikatu daitekeen.

- X aldagaiko balioak r klaseetan klasifikatzen dira eta Y aldagaikoak berriz c klaseetan. Izan bedi f_{ij} lehenengo ezaugarriko i klaserako eta aldi berean bigarren ezaugarriko j klaserako behatutako maiztasun absolutua.

- Modu honetan datuak *kontingentzia taula* izeneko bi sarrerako taula batean ordenatu daitezke.

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_c	X-ren maiztasunak
x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1c}	f_{x1}
$\vdots$						
x_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ic}	f_{xi}
$\vdots$						
x_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rc}	f_{xr}
Y-ren maiztasunak	f_{y1}	$\cdots$	f_{yj}	$\cdots$	f_{yc}	n

5.1 Taula. *r* x *c*-ko kontingentzia taula.

- Non $f_{xi} = \sum_{j=1}^{c} f_{ij}$; $f_{yj} = \sum_{i=1}^{r} f_{ij}$; $n = \sum_{i=1}^{r}\sum_{i=1}^{c} f_{ij}$.
- Izan bitez $p_{xi} = P(X = x_i)$, $p_{yj} = P(Y = y_j)$ eta $p_{ij} = P(X = x_i \cap Y = y_j)$.

- X eta Y aldagaiak independenteak balira $p_{ij} = P(X = x_i) \cdot P(Y = y_j) = p_{xi} \cdot p_{yj}$.
- Beraz ondorengo hipotesi kontrastea planteatzen da:

$$H_0 : p_{ij} = p_{xi} \cdot p_{yj} \quad \forall i = 1, 2, ..., r, \forall j = 1, 2, ..., c$$
$$H_a : \exists (i, j) \mid p_{ij} \neq p_{xi} \cdot p_{yj}$$

- Hipotesi nulua egia dela suposatuz, esperotako maiztasunak, e_{ij}, kalkulatuko dira. Horretarako, $\hat{p}_{xi} = \frac{f_{xi}}{n}$ eta $\hat{p}_{yj} = \frac{f_{yj}}{n}$ estatistikoak erabiliko dira. Horrela, $e_{ij} = n \cdot \hat{p}_{xi} \cdot \hat{p}_{yj} = n \cdot \frac{f_{xi}}{n} \cdot \frac{f_{yj}}{n} = \frac{f_{xi} \cdot f_{yj}}{n}$
- Hipotesi kontrasterako ondorengo estatistikoa erabiliko da:

$$\frac{\sum_{i=1}^{r} \sum_{j=1}^{c} (f_{ij} - e_{ij})^2}{e_{ij}}$$

- $e_{ij} \geq 5 \ \forall i = 1, 2, ..., r, \forall j = 1, 2, ..., c$ badira, kontrasterako estatistikoak $\chi^2_{(r-1)(c-1)}$ banaketa jarraituko luke.

- Beraz, independentzia probarako planteatutako hipotesi kontrasteari dagokion eskualde kritikoa α adierazgarritasun mailaz, ondorengoa litzateke:

$$EK_{\alpha} = \left[\chi^2_{(r-1)(c-1);\alpha}, +\infty \right)$$

- Bestalde, $(r-1)(c-1)=1$ denean, hau da, kontrasterako estatistikoak jarraituko lukeen banaketaren askatasun gradua 1 denean (2x2 kontingentzia tauletarako) eta $\exists (i,j) \left| e_{ij} < 5 \right.$, Yates-en jarraitasun zuzenketa aplikatu beharko litzateke. Horrela, kontrasterako estatistikoa honakoa litzateke:

$$\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\left| f_{ij} - e_{ij} \right| - 0.5\right)^2}{e_{ij}}$$

- Estatistiko honek $\chi^2_{(r-1)(c-1)=1}$ banaketa jarraituko luke.

Bi aldagaien independentziarako χ^2-ren proba R software librea erabiliz

- Independentzia probak egiteko R software-an kontingentzia taula idatzi behar da eta erabili beharreko komandoa *chisq.test* litzateke. Honela definituko da proba:

Bi aldagaien independentzia proba	
Komandoa	**Aukerak** (2x2-kontingentzia tauletan)
`chisq.test(Kontingentzia taula)`	`correct=T(Yates-en zuzenketa aplikatzeko) *Lehenespenez aukera hau dago.`
	`correct=F (Yates-en zuzenketa ez aplikatzeko)`

5.2 Taula. Independentzia probak egiteko erabiliko liratekeen R-ko komandoak eta aukera ezberdinak

- Independentzia proba egiterakoan R-k kontrasterako estatistikoaren balioa eta p-balioak adierazten ditu.

5.3. Normaltasun probak

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Datuen normaltasunaren analisia

- Datuen normaltasuna aztertzeko modu egokiena normaltasun probak dira. Proba hauetan ondorengo hipotesi kontrastea planteatzen da:

  H_0: Laginaren jatorria polupazio normala bat da.
  H_1: Laginaren jatorria ez da populazio normala bat.

- Ondoren Kolmogórov-Smirnov (K-S) eta Shapiro-Wilk (S-W) probak deskribatzen dira, non laginaren tamainaren arabera (K-S, $n > 50$ denean eta S-W, $n \leq 50$ denean) proba bat edo bestea aplikatuko den:

5.3.1. Kolmogórov-Smirnov

- K-S proba populazioaren normaltasun-hipotesia kontrastatzeko aplikatzen denean, probako estatistikoa diferentziarik (distantziarik) handiena da:

$$D = max\ |F_n(x) - F_0(x)|$$

non $F_n(x)$ laginaren banaketa funtzioa den, eta orokorrean, $F_0(x)$ berriz, hipotesi nuluan zehaztutako funtzio teorikoaren banaketa funtzioa.

- Normaltasun probaren kasuan, hipotesi nuluan zehaztutako funtzio teorikoa normala izango da.

- Orokorrean, K-S estatistikoa edo K-S-ren distantzia (D), bi laginen funtzio enpirikoen banaketen distantzia bertikal maximoa edo eta, funtzio enpiriko baten eta erreferentziazko funtzio teoriko baten banaketa funtzioen arteko distantzia maximoa litzateke.

- D estatistikoa banaketa funtzio metatuen arteko diferentziak aztertzeko sentikorra da. Sentikortasun hau lokalizaziorekikoa eta formarekikoa da.

- Laginaren jatorria populazio normal bat dela onartzeko, konparatzen diren funtzioen arteko distantzia D estatistikoa baino txikiagoa edo berdina izan behar du. p-balioa kalkulatzen da horretarako, eta balio hau zehaztutako adierazgarritasun maila baino handiago denean hipotesi nulua onartuko litzateke eta beraz, laginaren jatorria populazio normal batetik ez dela esanguratsuki aldentzen esan daiteke.

K-S proba R software librea erabiliz

- Normalean populazioaren parametroak ezagutzen ez direnez, K-S proba Lilliefors zuzenketa eginez burutzea gomendatzen da. Horretarako, R software-an lehenik *nortest* paketea instalatu eta kargatu behar da. Ondoren *lillie.test* komandoa erabiliko litzateke proba burutzeko.

```
> install.packages("nortest")
> library(nortest)
```

- Honela definituko da testa:

K-S normaltasun proba (Lilliefors zuzenketarekin) (n>50)	
Komandoa	**Argumentua**
`lillie.test()`	`lagina`

5.3.1 Taula. K-S test-a egiteko erabiliko liratekeen R-ko komandoa eta argumentua

K-S test-a egiterakoan, R-k, kontrasterako D estatistikoaren balioa eta p-balioa adierazten ditu.

5.3.2. Shapiro-Wilk

- S-W proba populazioaren normaltasun-hipotesia kontrastatzeko aplikatzen denean, probako estatistikoa ondorengoa da:

$$W = \frac{\sum_{i=1}^{n} (a_i \cdot x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

non

$x_{(i)}$ lagina ordenatu ondoren i-posizioan dagoen balioa den

a_i laginaren tamainaren arabera tabulaturiko koefizienteak diren (Flores et al., 2019)

- S-W testak zoriz hartzen den lagin bat banaketa normala duen populazio batetik etorri daitekeen aztertzen du.
- S-W probak W estatistikoaren balioa ematen digu eta orokorrean hau txikia denean ezin onar daiteke laginaren jatorria populazio normal bat denik.
- S-W testak limitazioak ditu, batez ere laginaren tamainarekiko: zenbat eta handiagoa izan laginaren tamaina emaitza esanguratsuagoa izango da estatistikoki.
- Laginaren jatorria populazio normal bat dela onartzeko, p-balioa kalkulatzen da horretarako eta balio hau zehaztutako adierazgarritasun maila baino handiago denean hipotesi nulua onartuko litzateke; beraz, laginaren jatorria populazio normal batetik ez dela esanguratsuki aldentzen esan daiteke.

S-W proba R software librea erabiliz

- Normaltasun probetan laginaren tamaina 50 edo txikiagoa denean, K-S proba erabili beharrean Shapiro-Wilk proba erabiltzen da. Horretarako, R software-an *shapiro.test* komandoa erabiliko litzateke proba burutzeko.

- Honela definituko da testa:

S-W normaltasun proba ($n \leq 50$)	
Komandoa	**Argumentua**
`shapiro.test()`	`lagina`

5.3.2 Taula. *S*-W test-a egiteko erabiliko liratekeen R-ko komandoa eta argumentua

S-W testa egiterakoan, R-k kontrasterako W estatistikoaren balioa eta p-balioa adierazten ditu.