Tema 4. Contrastos d'hipòtesis NO-paramètriques

4.0. Introducció

Fins ara es realitzàvem contrastos sobre el valors assignats a alguns paràmetres (conjunt de paràmetres) que definien una determinada funció de probabilitat f(X;θ). Això implicava:

1. La mostra obtinguda procedia d'una població amb una distribució CONEGUDA.

2. Els procediments d'inferència es centraven en paràmetres DESCONEGUTS d'aquesta funció.

3. S'exigia robustesa a la tècnica inferencial per a que les vulneracions de les hipòtesis no afectessin als CONTRASTOS o a les ESTIMACIONS.

CONTRAST NO-PARAMÈTRIC: Es desenvolupen per a evitar aquest conjunt d'hipòtesis i que es contamini el procés davant variacions dels mateixos.

DUES TÈCNIQUES NO-PARAMÈTRIQUES:

1. El procés es basa en un ESTADÍSTIC de CONTRAST no vinculat amb cap paràmetre poblacional.

2. METODES DE DISTRIBUCIÓ LLIURE. L'ESTADÍSTIC presenta una distribució que NO DEPÈN de la DISTRIBCIÓ DE PROBABILITAT de la POBLACIÓ.

• Generalment no es fan servir els VALORS DE LA MOSTRA, sinó la seva FREQÜÈNCIA i/o el seu ORDRE (en diferents formes).

• Són els únics aplicables si les observacions són NOMINALS/ORDINALS

• Són més EFICIENTS que els seus HOMÒLEGS si la població és NO-NORMAL.

Tema 4. Contrastos d'hipòtesis NO-paramètriques

Contrastos no-paramètrics més rellevants:

Mostra	**Escala Nominal**	**Escala Ordinal**
Una mostra	**X^2** **Ratxes** Binomial	Kolmogorov-Smirnov
Dues mostres relacionades	Mcnemar	Signes **Wilcoxon**
K mostres relacionades	Q de Cochran	Friedman Kendall
Dues mostres independents		Mediana Mann-Whitney Kolmogorov-Smirnov Wald-Wolfowitz Moses
K mostres independents		Mediana Kruskal-Wallis

Contrastos específics de Normalitat:

Shapiro-Wilk
Lilliefors
Jarque-Bera
Dornik-Hansen

Contrastos de relació entre atributs:

Taules de contingència:

Independència
Homogeneïtat

(*) En negreta els que s'analitzaran aquest curs.

Tema 4. Contrastos d'hipòtesis NO-paramètriques

4.1. Una mostra. Bondat de l'Ajust

OBJECTIU: Verificar si una mostra procedeix d'una població amb una DETERMINADA DISTRIBUCIÓ DE PROBABILITAT.

$$\begin{cases} H_0 :"L'Ajust\ és\ bo" \\ H_1 :"No\ H_0" \end{cases}$$

Test χ^2: Contrastar si una mostra procedeix d'una determinada distribució de probabilitat

CAS 1: Es suposa que F(X) està COMPLETAMENT ESPECIFICADA → NO EXISTEIXEN PARÀMETRES DESCONEGUTS

$$\begin{cases} H_0 : F(X) = F_0(X) \\ H_1 : F(X) \neq F_0(X) \end{cases}$$

CAS 2: Es suposa que F(X) NO ESTÀ COMPLETAMENT ESPECIFICADA → ESTIMAR PARÀMETRES DESCONEGUTS

CAS 1: Es disposa d'una mostra $X=\{X_1, \ldots..,X_N\}$ ALEATÒRIA.

Es classifiquen les N observacions en "r" classes MUTUAMENT EXCLOUENTS de forma que:

$$\chi = A_1 \cup A_2 \cup \cdots \cup A_r$$ → Tot el camp de variació mostral

Tema 4. Contrastos d'hipòtesis NO-paramètriques

Clase A_i conté n_i elements $\rightarrow N = \sum_{i=1}^{r} n_i \rightarrow$ Mida mostral

Sota el supòsit que H_0 és certa → $\begin{cases} F(A_i) = p_i \\ \sum_{i=1}^{r} p_i = 1 \end{cases}$

Al procedir d'una MAS la:

Probabilitat d'aparèixer → $n_1 \in A_1$
$\vdots$
Probabilitat d'aparèixer → $n_i \in A_i$
$\vdots$
Probabilitat d'aparèixer → $n_r \in A_r$

→ Segueix una distribució MULTINOMIAL com la següent:

$$P(n_1, n_2, ..., n_r) = \frac{N!}{n_1! n_2! \cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$

On cada:

$$n_i \approx B(n, p_i) \rightarrow E(n_i) = np_i = \underbrace{E_i}_{\text{valor esperat}}$$

Per tant, cada E_i és el VALOR ESPERAT del nombre d'observacions que pertanyen a la classe A_i

Tema 4. Contrastos d'hipòtesis NO-paramètriques

Degut a la aleatorietat del mostreig, és d'esperar que les FREQÜÈNCIES OBSERVADES no siguin les mateixes que les FREQÜÈNCIES ESPERADES. Per tant, s'ha de valorar ESTADÍSTICAMENT si aquestes DISCREPÀNCIES són SUFICIENTMENT IMPORTANTS O NO. Així, l'ESTADÍSTIC del contrast és:

$$\sum_{i=1}^{r}\frac{(n_i-E_i)^2}{E_i}=\sum_{i=1}^{r}\frac{(n_i-Np_i)^2}{Np_i}\approx\chi^2_{r-1}$$

Si les DISCREPÀNCIES són conjuntament GRANS → Refusem H_0 → La informació mostral indica que no hi ha evidència per a creure que la mostra procedeix d'una població amb la distribució de probabilitat com la proposada per la H_0

$$P\left(\sum_{i=1}^{r}\frac{(n_i-Np_i)^2}{Np_i}\geq\chi^2_{r-1,1-\alpha}\right)=\alpha$$

NOTES: 1º.- És comporta com un χ^2 assimptòticament → $\forall E_i\geq 5$ en cas contrari reagrupar categories. 2º.- Es pot aplicar tant a variables DISCRETES (incloses ordinals i nominals) I CONTINUES. 3º.- Es aconsellable que la construcció de les classes presenti una probabilitat semblant.

ECET 4.1/4.2

Tema 4. Contrastos d'hipòtesis NO-paramètriques

CAS 2: S'ha d'estimar els paràmetres:

$$\cancel{P(A_i)=p_i}$$

$$P(A_i;\theta_1,\theta_2,\cdots,\theta_k)=p_i(\theta_1,\theta_2,\cdots,\theta_k)$$

$$\sum_{i=1}^{r}\frac{(n_i-\hat{E}_i)^2}{\hat{E}_i}=\sum_{i=1}^{r}\frac{(n_i-Np_i(\theta_1,\theta_2,\cdots,\theta_k))^2}{Np_i(\theta_1,\theta_2,\cdots,\theta_k)}\approx\chi^2_{r-k-1}$$

$$P\left(\sum_{i=1}^{r}\frac{(n_i-\hat{E}_i)^2}{\hat{E}_i}\geq\chi^2_{r-k-1,1-\alpha}\right)=\alpha$$

On *K* és el nombre de paràmetres que s'han d'estimar i, per tant, implica una pèrdua de graus de llibertat.

ECET 4.3

Tema 4. Contrastos d'hipòtesis NO-paramètriques

4.2. Una mostra. Test de Normalitat

Shapiro-Wilk

$$\begin{cases} H_0 :"És\ Normal" \\ H_1 :"No\ H_0" \end{cases}$$

OBJECTIU: És un contrast específic per a verificar si una mostra procedeix d'una distribució normal sense haver de fer cap especificació sobre els seus paràmetres. És molt útil amb mostres de mida petita n<50. Es disposa d'una mostra X={X_1, …..,X_N} ALEATÒRIA. Per a realitzar el contrast:

$$\begin{cases} H_0 :"X\ és\ distribueix\ com\ una\ Normal" \\ H_1 :"No\ H_0" \end{cases}$$

1º.- Es fa l'ordenació de la mostra de menor a major

$$u_1 \le u_2 \le \cdots \le u_N$$

2º.- S'obtenen els següents estadístics:

$$\bar{u} = \frac{\sum_{i=1}^{N} u_i}{N} \quad i \quad z^2 = \sum_{i=1}^{N} (u_i - \bar{u})^2$$

3º.- Es computa el següent coeficient:

$b = \sum_{i=1}^{k} a_{N-i+1}(u_{N-i+1} - u_i)$ on "a" són els coeficients tabulats per Shapiro-Wilk, 1965

$k = \frac{N}{2}$ si n és parell i $k = \frac{N-1}{2}$ si n és senar

4º.- L'estadístic del contrast és:

$W = \frac{b^2}{z^2}$ Si $P(W \le "\text{Valor taules W}") = \alpha \rightarrow$ es RH_0

ECET 4.4

Software estadístic: Gretl, R, SPSS, EViews, etc..

Tema 4. Contrastos d'hipòtesis NO-paramètriques

4.2. Una mostra. Aleatorietat

OBJECTIU: Verificar si un conjunt d'observacions constitueixen una mostra aleatòria procedent d'una població CONTÍNUA

$$\begin{cases} H_0 :"És\ aleatòria" \\ H_1 :"No\ H_0" \end{cases}$$

Test de Ratxes: Una ratxa és una succió de símbols de la mateixa classe que es troba limitat per símbols d'una altra classe.

Exemple 1: QQQQHHHHQQHHHQQQQQQHH

Amb aquests dos tipus d'observacions (Q i H) diem que hi ha 6 ratxes, 3 de Q i 3 d'H

Exemple 2: QQQQQQQQQQHHHHHHHHHHH

Exemple 3: QHQHQHQHQHQHQHQHQHQ

La idea és que a l'exemple 1 podríem dir que si les observacions procedeixen d'una mateixa població, els símbols Q i H apareixeran barrejats i sense cap mena de sistematització. En canvi, als exemples 2 i 3 ens trobem amb dos sistematitzacions extremes. El cas 2, que presenta molt poques ratxes (2) i, el cas tres, moltes ratxes (21), tantes com elements.

Per tant, el test de ratxes pretén distingir si ens trobem amb un cas com l'exemple 1, aleatori, o com els altres dos, sistemàtics.

Tema 4. Contrastos d'hipòtesis NO-paramètriques

Per a instrumentalitzar les ratxes:

Variables continues amb un atribut dicotòmic de referència (edat, puntuació, etc) de (home/dona, matí/tarda, GrupaA/GrupB, etc..)

1º.- S'obtenen les ratxes.
2º.- Si el nombre qualsevol de ratxes és inferior a 20. **(Taules específiques: p.ex.: Casas Sánchez, et al. 2006, pags 501-502)**
3º.- Per valors superiors el nombre de ratxes, R, es distribueix com una normal amb paràmetres:

$$R \approx N\left(E(R); \sqrt{VAR(R)}\right)$$

$$E(R) = \frac{2N_+N_- + n}{n}$$

$$VAR(R) = \frac{2N_+N_-(2N_+N_- - n)}{n^2(n-1)}$$

Contrast a doble cua i l'estadístic del contrast s'instrumentalitza com qualsevol contrast basat en una distribució normal.

ECET 4.5

On N_+ i N_- són el nombre "d'elements" de les ratxes positives i negatives $N_+ + N_- = n$

Variables continues sense atribut dicotòmic de referència (PIB, edat, puntuació):
Es realitza exactament igual que en el cas anterior només que per a obtenir les ratxes es fa la diferència entre el valor de la variable i la mediana. Després s'eliminen els valors zero reduint la mostra. Finalment, els valors negatius determinaran una ratxa i els positius l'altre.
(Software estadístic: Gretl, Spss, Eviews, R, etc..)

Tema 4. Contrastos d'hipòtesis NO-paramètriques

4.3. Dues mostres. Dades emparellades

$$\begin{cases} H_0 & Me_X = Me_Y \\ H_1 & Me_X \neq Me_Y \end{cases}$$

Test de Wilcoxon

OBJECTIU: Pertany al conjunt de test que es denominen de localització i pretén verificar si un conjunt d'observacions $\{X_i, Y_i\}$ d'una distribució BIDIMENSIONAL presenten la mateixa mesura de posició o localització. En aquest cas la mediana. Aquest test és un exemple, ni han d'altres metodologies i per altres mesures de localització/posició. És útil per a valorar la situació de dos individus després que hi hagi passat en el temps algun fenomen. Ex.(valor de l'individu abans d'un tractament i valor del mateix individu després del tractament)

Per a instrumentalitzar el contrast es realitza el següent:

ECET 4.6

1º.- Es calcula les diferències entre les variables $d_i = X_i - Y_i$

2º.- S'eliminen les diferències nul·les

3º.- S'assignen els rangs a aquestes diferències de menor a major. (En cas d'empat, s'assigna el rang mitjà).

4º.- Anomenem R^+ a la suma dels rangs que presenten les diferències positives i R^- a la suma dels rangs que presenten les diferències negatives. Es pot demostrar que:

$$R^+ + R^- = \frac{n(n+1)}{2}$$

5º.- Existeixen valors tabulats per exemple a Casas-Sánchez 2006, pag 500. Ara bé, si la mostra és suficientment gran es compleix que:

$$R^+ \approx N\left(\frac{n(n+1)}{4}; \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$$

Contrast a doble cua i l'estadístic del contrast s'instrumentalitza com qualsevol contrast basat en una distribució normal.

(Software estadístic: Gretl, Spss, Eviews, R, etc..)

Tema 4. Contrastos d'hipòtesis NO-paramètriques

4.3. Taules de contingència: Test d'independència

OBJECTIU: Verificar la independència entre dos factors

$$\begin{cases} H_0 : "Són\ Independents" \\ H_1 : "No\ H_0" \end{cases}$$

	B_1	B_2	.....	B_j	.....	B_c	*Freqüències marginals*
A_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1c}	$n_{1.}$
A_2	n_{21}	n_{22}	.....	n_{2j}	.....	n_{2c}	$n_{2.}$
.....	.....	.....	.....	.....	.....	.....	.....
A_i	n_{i1}	n_{i2}	.....	n_{ij}	.....	n_{ic}	$n_{i.}$
.....	.....	.....	.....	.....	.....	.....	.....
A_r	n_{rc}	n_{rc}	.....	n_{rj}	.....	n_{rc}	$n_{r.}$
Freqüències marginals	$n_{\cdot 1}$	$n_{\cdot 2}$	.....	$n_{\cdot j}$	.....	$n_{\cdot c}$	n

$$\sum_{i=1}^{r} n_{ij} = n_{\bullet j} \qquad \sum_{j=1}^{c} n_{ij} = n_{i\bullet} \qquad \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = n$$

Tema 4. Contrastos d'hipòtesis NO-paramètriques

Condició d'independència (**Estadística Econòmica i Empresarial I**)

Dos variables són INDEPENDENTS si la seva funció de distribució de probabilitat conjunta és igual al producte de les marginals. Per tant, sota la H_0

$$P_{ij} = P(A_i \cap B_j) = P(A_i) * P(B_i) = P_{i\bullet} * P_{\bullet j} \quad \forall i = 1 \cdots r, i\ \forall j = 1 \cdots c$$

RECORDATORI. Només amb que una parella de punts no fos així, ja no hi hauria independència.

Distribució de probabilitat conjunta:

$P(A_i \cap B_j)$	$P(A_i) * P(B_i)$
P_{11}	$P_1 * P_1$
$\vdots$	$\vdots$
P_{ij}	$P_i * P_j$
$\vdots$	$\vdots$
P_{rc}	$P_r * P_c$

← Totes aquestes probabilitats han de ser estimades a partir de la informació mostral

Tema 4. Contrastos d'hipòtesis NO-paramètriques

	B_1	B_2	.....	B_j	.....	B_c	*Freqüències marginals*
A_1	P_{11}	P_{12}	.....	P_{1j}	.....	P_{1c}	$P_{1.}$
A_2	P_{21}	P_{22}	.....	P_{2j}	.....	P_{2c}	$P_{2.}$
.....	.....	.....	.....	.....	.....	.....	.....
A_i	P_{i1}	P_{i2}	.....	P_{ij}	.....	P_{ic}	$P_{i.}$
.....	.....	.....	.....	.....	.....	.....	.....
A_r	P_{rc}	P_{rc}	.....	P_{rj}	.....	P_{rc}	$P_{r.}$
Freqüències marginals	$P_{\cdot 1}$	$P_{\cdot 2}$	.....	$P_{\cdot j}$	.....	$P_{\cdot c}$	1

Per tant, s'han d'estimar r+c-2 paràmetres. El -2 és pel fet que dos paràmetres es poden extreure per la diferència amb la unitat, atès que: $\sum_{i=1}^{r} P_{i\bullet} = 1 \quad \sum_{j=1}^{c} P_{\bullet j} = 1$

Fem servir el MÈTODE DE LAGRANGIÀ

1º.– Obtenim la funció de Versemblança: $\ln\left[L\left(P_{i\bullet}, P_{\bullet j}\right)\right] = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \ln\left(P_{i\bullet}, P_{\bullet j}\right)$

2º.- Subjecte a dues restriccions: $\sum_{i=1}^{r} P_{i\bullet} = 1 \quad i \quad \sum_{j=1}^{c} P_{\bullet j} = 1$

3º.– Obtenim els ESTIMADORS MÀXIMVERSEMBLANTS: $\hat{P}_{i\bullet} = \frac{n_{i\bullet}}{n} \quad i \quad \hat{P}_{\bullet j} = \frac{n_{\bullet j}}{n}$

Tema 4. Contrastos d'hipòtesis NO-paramètriques

Contrast:

1º.- Sota la H_0 $\quad P_{ij} = P\left(A_i \cap B_j\right) = P\left(A_i\right)P\left(B_i\right) = P_{i\bullet}P_{\bullet j} \quad \rightarrow \quad \hat{P}_{ij} = \hat{P}_{i\bullet}\hat{P}_{\bullet j}$

Per tant, obtindrem els VALORS ESPERATS per cada creuament que hauríem d'esperar en el cas que les dues variables FOSIN INDEPENDENTS. Aquests són:

$$\hat{E}_{ij} = n\hat{P}_{ij} = n\hat{P}_{i\bullet}\hat{P}_{\bullet j} = \frac{n_{i\bullet}n_{\bullet j}}{n}$$

Així, si les freqüències estimades són MOLT DIFERENTS de les freqüències observades, refusem H_0

2º.- L'estadístic del contrast és:

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}} \approx \chi^2_{g.d.ll?????}$$

Quants graus de llibertat??. g.d.ll=total categories - nº paràmetres a estimar - 1

g.d.ll=(r*c)-(r+c")-1=(r-1)*(c-1)

$$Si \quad P\left(\sum_{j=1}^{c} \frac{\left(n_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}} \geq \chi^2_{(r-1)(c-1)}\right) = \alpha \quad \rightarrow \quad RH_0$$

ECET 4.7