Sobre la determinació numèrica de la fidelitat en bases de dade

-

Quadre 2.2.2 : Estadístics per a avaluar la fidelitat d'un tàxon a una unitat de vegetació o sintàxon a partir de presències i absències, i les seves relacions. Modificat a partir de Chytrý et al. (2002).

Siguin:
Ω_A un subconjunt d'inventaris relatius a un sintàxon o unitat de vegetació.
N el nombre d'objectes o inventaris totals
N_A el nombre d'objectes o inventaris d' Ω_A.
n el nombre d'aparicions del tàxon en el conjunt de dades.
n_A el nombre d'aparicions del tàxon en Ω_A.

Es poden definir els següents estadístics per avaluar la fidelitat d'un tàxon al conjunt Ω_A:

1. u hipergeomètrica (Bruelheide 2000, Bruelheide & Chytrý 2000):

2. u binomial (Bruelheide 1995,2000; Bruelheide & Chytrý 2000):

3. Chi quadrat (χ^2, Sokal & Rohlf 1995: 736):
A partir de les caselles d'una taula de contingència 2x2 és habitual construir l'estadístic:

Podem traduïr les caselles de la taula: , , , i . Substituint, arribem a una expressió de χ^2 en la mateixa notació que els anteriors:

4. Coeficient phi (Φ, Sokal & Rohlf 1995:741-743):

La significació dels estadístics Φ i u_{hyp}.

El coeficient Φ és l'únic dels estadístics del quadre 2.2.2 que pren valors acotats: $\Phi \in [-1, +1]$, on $+1$ denota una preferència total, 0 denota indiferència, i -1 denota que el tàxon apareix arreu excepte a la unitat d'interès. En realitat, es tracta d'una mesura de correlació entre variables binàries, essent equivalent a la correlació r de Pearson quan els valors de les variables són binaris. A la taula 2.2.2 hem calculat els valors de Φ per a cada una de les combinacions de la

classe de presència al grup d'interès (0, I, ... V) i la classe de presència a la resta d' inventaris (G0, GI, ... GV). Si ens fixem en la primera columna de la taula, entendrem millor la interacció de la selectivitat i grau de presència com a factors de la fidelitat: Per a un tàxon totalment absent als inventaris externs al sintàxon (columna G0), el valor de fidelitat serà més alt depenent de la classe de presència del tàxon dins del sintàxon, mentre que la selectivitat, exceptuant la primera casella, és sempre màxima.

	G0	GI	GII	GIII	GIV	GV
0	N/A	-0.16	-0.25	-0.36	-0.53	-1.00
I	0.43	0	-0.12	-0.24	-0.41	-0.88
II	0.61	0.14	0	-0.12	-0.28	-0.76
III	0.76	0.28	0.12	0	-0.14	-0.61
IV	0.88	0.41	0.24	0.12	0	-0.43
V	1.00	0.53	0.36	0.25	0.16	0

Taula 2.2.2: Valors del coeficient **Φ** de diferents combinacions de presència al sintàxon d'interès (0-V) i a la resta d'inventaris externs al sintàxon (G0-GV). La proporció d'inventaris del sintàxon d'estudi respecte el total d'inventaris és $P_A = N_A/N = 0.5$.

El valor del coeficient Φ és independent del nombre d'objectes total (N), mentre que la resta d'estadístics relacionats al quadre 2.2.2 augmenten de valor en augmentar N. Aquesta és la raó per la qual Φ és l'estadístic més adequat per comparar fidelitats calculades en conjunts de dades de mida diferent (Chytrý *et al.* 2002). Per contra, aquesta avantatge és a la vegada un inconvenient, ja que el coeficient Φ, a diferència dels altres, no conté informació de significació estadística. És a dir, que valors més alts de l'estadístic no impliquen més significació de la fidelitat. Per a testar la significació de Φ hom pot fàcilment multiplicar el valor pel factor $\sqrt{(N-1)}$. Amb aquesta transformació s'obté l'estadístic u_{hyp}, el qual sí que ens informa de la significació estadística de la fidelitat. Com que la distribució hipergeomètrica d' u_{hyp} es pot aproximar a una distribució normal estandarditzada, el seu valor es pot descriure com el nombre de desviacions estàndard en que n_A es distancia del valor que esperaríem si el tàxon i la unitat de vegetació fossin independents. Així, hom pot afirmar que valors $|u_{hyp}| = |\Phi \cdot \sqrt{(N-1)}| > 1.96$ són estadísticament significatius amb una probabilitat d'error $\alpha < 0.05$ (Chytrý *et al*. 2002). Si hom desitja establir un llindar de fidelitat per al coeficient Φ, que anomenarem Φ_t, es pot saber quina N és la mínima necessària per tal de tenir significació estadística. Senzillament:

$$N_t(\alpha, \Phi_t) = \left(\frac{u_{hyp}(\alpha)}{\Phi_t}\right)^2 + 1 = \left(\frac{u_{hyp}(\alpha)}{\Phi_t}\right)^2 + 1$$

Per exemple, si establim que $\Phi_t = 0.3$, llavors $N_t(\alpha = 0.05, \Phi_t = 0.3) \approx 46$ i $N_t(\alpha = 0.01, \Phi_t = 0.3) \approx 61$. Per tant, si emprem més de 61 inventaris per determinar tàxons diagnòstics en base a aquest llindar de fidelitat, els tàxons trobats seran significativament diagnòstics amb una probabilitat d'error $\alpha < 0.01$.

A banda de l'aproximació a la normal estandarditzada, hom pot emprar el test exacte de Fisher per a conèixer la significació d'aquests estadístics de la fidelitat. Concretament, hom calcula la probabilitat de trobar valors més extrems de n_A que l'observat mitjançant:

$$P(n_A \geq \hat{n}_A) = \sum_{i=1}^{\hat{n}_A} \frac{n! \cdot N_A! \cdot (N - n)! \cdot (N - N_A)!}{i! \cdot N! \cdot (n - i)! \cdot (N_A - i)! \cdot (N - N_A - n + i)!}$$

En darrer lloc, hem de dir que hom no pot confiar excessivament en la significació estadística de valors de fidelitat de tàxons a sintàxons, car és fàcil augmentar la significació incrementant N. És a dir, podem augmentar la significació introduint nous inventaris on hom pot esperar que el tàxon sigui rar o absent. Desenvoluparem més aquest raonament en els propers apartats.