Problem in parametrični prostor

Verjetnostni prostor Ω razbijemo na disjunktni uniji:

- $\Omega = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_r$ in
- $\Omega = B_1 \sqcup B_2 \sqcup \cdots \sqcup B_s$.

Privzamemo, da za vsa števila i in j velja $p_{ij} = P(A_i \cap B_j) > 0$. (Sledi tudi $p_i = P(A_i) > 0$ in $q_j = P(B_j) > 0$.) Velja torej

$$\Theta = \left\{ [p_{ij}]_{1 \leqslant i \leqslant r, 1 \leqslant j \leqslant s} \;\middle|\; p_{ij} > 0, \sum_{i,j} p_{ij} = 1 \right\},$$

kar je simpleks razsežnosti $rs - 1$, ki naravno leži v prostoru matrik $\mathbb{R}^{r \times s}$.

Ničelna hipoteza neodvisnosti

Neodvisnost: $\forall i, j : \; p_{ij} = P(A_i \cap B_j) = P(A_i) \cdot P(B_j) = p_i \cdot q_j$.

To pomeni, da za vsaki števili i in j velja $p_{ij} = \left(\sum_k p_{ik}\right) \cdot \left(\sum_l p_{lj}\right)$.

Ker je ničelna hipoteza očitno parametrizirana s kartezičnim produktom $\mathbb{\Delta}^{r-1} \times \mathbb{\Delta}^{s-1}$, je $\dim H_0 = r + s - 2$. Velja

$$\dim\Theta - \dim H_0 = rs - 1 - (r + s) - 2 = rs - r - s + 1 = (r-1) \cdot (s-1).$$

Opis eksperimenta oziroma vzorčenja

Neodvisno in slučajno izberemo n elementov iz množice Ω in s T_{ij} označimo število elementov, ki pripadajo preseku $A_i \cap B_j$.
Rezultate predstavimo v kontingenčni tabeli:

	1	2	$\dots$	$s-1$	s	
1	T_{11}	T_{12}	$\dots$	$T_{1,s-1}$	T_{1s}	U_1
2	T_{21}	T_{22}	$\dots$	$T_{2,s-1}$	T_{2s}	U_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
r	T_{r1}	T_{r2}	$\dots$	$T_{r,s-1}$	T_{rs}	U_r
	V_1	V_2	$\dots$	V_{s-1}	V_s	n

Gre seveda za diskretno slučajno spremenljivko X s končno zalogo vrednosti $\{(i,j) \mid 1 \leqslant i \leqslant r,\ 1 \leqslant j \leqslant s\}$, kjer je $X(\omega) = (i,j)$, če $\omega \in A_i \cap B_j$.

Testiranje neodvisnosti: razmerje verjetij

Funkcija verjetja za vzorec velikosti n in cenilka največjega verjetja

Kot prej: $L(x_1, \ldots, x_n; \mathbf{p}) = \prod_{i,j} p_{ij}^{T_{ij}}$,

kjer je T_{ij} število parov (i, j) med vrednostmi $x_1, \ldots, x_n$.

CNV je $\hat{p}_{ij} = \frac{T_{ij}}{n}$, zato: $\sup_{\mathbf{p}\in\Theta} L(\mathbf{x}; \mathbf{p}) = \prod_{ij} \left(\frac{T_{ij}}{n}\right)^{T_{ij}}$.

Cenilka največjega verjetja na H_0

Upoštevamo, da velja $p_{ij} = p_i q_j$ in računamo:

$$L(\mathbf{x}; [p_i q_j]_{ij}) = \prod_{ij} (p_i q_j)^{T_{ij}} = \prod_{ij} p_i^{T_{ij}} q_j^{T_{ij}} = \left(\prod_{ij} p_i^{T_{ij}}\right) \cdot \left(\prod_{ij} q_j^{T_{ij}}\right)$$

$$= \left(\prod_i p_i^{\sum_j T_{ij}}\right) \cdot \left(\prod_j q_j^{\sum_i T_{ij}}\right) = \left(\prod_i p_i^{U_i}\right) \cdot \left(\prod_j q_j^{V_j}\right).$$

Upoštevamo $\sum_i p_i = 1$ in $\sum_j q_j = 1$ in dobimo $\hat{p}_i = \frac{U_i}{n}$, $\hat{q}_j = \frac{V_j}{n}$.

Torej: $\sup_{\mathbf{p}\in H_0} L(\mathbf{x}; \mathbf{p}) = \left(\prod_i \left(\frac{U_i}{n}\right)^{U_i}\right) \cdot \left(\prod_j \left(\frac{V_j}{n}\right)^{V_j}\right) = \cdots = \prod_{ij} \left(\frac{U_i V_j}{n^2}\right)^{T_{ij}}$.

Testna statistika

- $\lambda_n(\mathbf{x}) = \prod_{ij} \left(\frac{U_i V_j}{n T_{ij}}\right)^{T_{ij}}$,
- $-2 \log \lambda_n(\mathbf{x}) = -2 \sum_{ij} T_{ij} \log \left(\frac{U_i V_j}{n T_{ij}}\right)$.

Test 1

Zavrnemo hipotezo neodvisnosti $\mathrm{H}_0 : p_{ij} = p_i \cdot q_j$, če $-2 \log \lambda_n(\mathbf{x}) > C$. Pri tem

- konstanto C določimo eksaktno z multinomsko porazdelitvijo za majhne vzorce ali
- vzamemo $C = \chi^2_{(r-1)(s-1);\alpha}$ za velike vzorce, kjer upoštevamo $-2 \log \lambda_n \xrightarrow[n\to\infty]{d} \chi^2_{(r-1)(s-1)}$.

Standardna χ^2-testa

Kot prej se izkaže, da z uporabo Taylorjevega polinoma 2. stopnje za logaritem dobimo dve statistiki, ki ravno tako v porazdelitvi konvergirata k porazdelitvi $\chi^2_{(r-1)(s-1)}$. To sta

- $\sum_{i,j} \frac{(T_{ij}-U_i V_j/n)^2}{T_{ij}}$ in
- $\sum_{i,j} \frac{(T_{ij}-U_i V_j/n)^2}{U_i V_j/n}$.

Običajni zapis

Pišimo $\hat{T}_{ij} = U_i V_j/n$. Števila $\hat{T}_{ij}$ imenujemo (glede na H_0) pričakovane frekvence, števila T_{ij} pa opažene frekvence.

$$\mathrm{H}_0 \text{ zavrnemo, če } \sum_{1\leqslant i\leqslant r, 1\leqslant j\leqslant s} \frac{(T_{ij}-\hat{T}_{ij})^2}{\hat{T}_{ij}} > \chi^2_{(r-1)(s-1);\alpha}.$$