Mesures de distance

Variables nominales

Véronique Tremblay

©Véronique Tremblay 2021

Variables nominales

$$d(i,j) = \sum_{k=1}^{K} \mathcal{J}(x_{ik} \neq x_{jk})$$

Exemple pour le calcul de variables binaires

Individu	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
i	1	0	0	0	0	1	0	0	0	0
j	1	0	0	0	0	0	1	0	0	0

La similarité est définie par

$$s(i,j) = 1 - \frac{d(i,j)}{K}$$

©Véronique Tremblay 2021

Variables asymétriques

Situations pour lesquelles deux individus qui ont une certaine caractéristique se ressemblent plus que deux individus qui ne l'ont pas.

©Véronique Tremblay 2021

Indice de Jaccard

Exemple pour le calcul de variables asymétriques

Région	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
i	1	0	0	0	0	1	0	0	0	0
j	1	0	0	0	0	0	1	0	0	0

$I = \{E1, E6\}$

$J = \{E1, E7\}$

$J(i,j) = \frac{|I \cap J|}{|I \cup J|}$

©Véronique Tremblay 2021

Indice de Jaccard

- On assigne la modalité 1 à la valeur la plus «rare».
- Indice de Jaccard (1901)

$$J(i,j) = \frac{\sum_{k=1}^{K} x_{ik}x_{jk}}{\sum_{k=1}^{K}\{1-(1-x_{ik})(1-x_{jk})\}}$$

©Véronique Tremblay 2021

Résumé

- Mesure de distance pour deux variables nominales
- Variables asymétriques

©Véronique Tremblay 2021

Références

Jaccard, Paul. 1901. “Étude Comparative de La Distribution Florale Dans Une Portion Des Alpes et Des Jura.” *Bulletin de La Société Vaudoise Des Sciences Naturelles* XXXV: 547–79.

©Véronique Tremblay 2021