

MÉTHODE DES RANGS ALÉATOIRES (C12, F6)

(24 / 10 / 2019, © Monfort, Dicostat2005, 2005-2019)

La **méthode des rangs aléatoires** permet de traiter le **problème des rangs multiples**, ie le cas où plusieurs **observations** sont égales, ie prennent une même valeur (cf **rang**, **rang multiple**). En général, les formules d'**inférence statistique** (eg **Statistique non paramétrique**), valables lorsque les observations sont toutes différentes deux à deux (**lois** à **densités** continues), doivent être « corrigées » lorsque des valeurs multiples sont observées (eg **lois discrètes**, calculs avec nombres décimaux).

(i) Soit $X = (X_1, \dots, X_N)$ un **échantillon** constitué de **vars** X_n , $X^{(.)}$ la **statistique d'ordre** (croissant) associée à X et R sa **statistique de rang**.

Avant la mise en oeuvre d'une **procédure statistique**, le procédé de « **perturbation** » **des valeurs initiales** consiste à remplacer une **séquence** de valeurs multiples quelconque tq :

$$(1) \quad X_n = \dots = X_{n+L} \quad (\text{où } n \in N_N^* \text{ et } 1 \leq L \leq N)$$

par une séquence de valeurs distinctes deux à deux, ie par :

$$(2) \quad X_n + \varepsilon_n, \dots, X_{n+L} + \varepsilon_{n+L},$$

où les ε_{n+l} ($l = 0, 1, \dots, L$) sont positifs et distincts deux à deux : eg $\varepsilon = (\varepsilon_n, \dots, \varepsilon_{n+L}) \in (\mathbf{R}_+^*)^{1+L} = \{\varepsilon \in (\mathbf{R}_+^*)^{1+L} : \varepsilon_n < \dots < \varepsilon_{n+L}\}$.

Dans une seconde étape, ie après procédure, on peut calculer les limites des **statistiques** d'intérêt (**estimateur**, **statistique de test**, etc) lorsque les nombres ε_{n+l} ($l = 0, 1, \dots, L$) tendent (indépendamment) vers zéro.

Cependant, une **singularité** apparaît souvent : non convergence, valeurs multiples.

(ii) La **méthode des rangs aléatoires** est fondée sur le principe suivant. On associe à X un échantillon fictif (ou **échantillon artificiel**) $U = (U_1, \dots, U_N)$, tiré selon la **loi uniforme** (continue) $\mathcal{U}(0, 1)$: dans ce cas, aucune coordonnée ne peut (théoriquement) être égale λ -p.p. une autre.

On définit alors une **relation d'ordre** tq (cf **ordre lexicographique**) :

$$(3) \quad (X_\alpha, U_\alpha) \prec (X_\beta, U_\beta) \Leftrightarrow \begin{array}{l} \text{ou bien } X_\alpha < X_\beta, \\ \text{ou bien } X_\alpha = X_\beta \text{ et } U_\alpha < U_\beta. \end{array}$$

La relation \prec permet ainsi d'ordonner « au hasard » l'échantillon associé $((X_1, U_1), \dots, (X_N, U_N))$ et l'on attribue alors à X_α le **rang** du couple (X_α, U_α) .