VARIABLE DE CORNFIELD (C1, M3)
(09 / 10 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion de **variable de CORNFIELD** intervient notamment en **théorie des sondages**, eg dans le cas d'un **sondage exhaustif**.

(i) Soit $\Omega = \{\omega_1 ,..., \omega_M\}$ une **population** finie et $A = \{a_1 ,..., a_N\}$ un **échantillon** de taille N extrait de Ω selon un **plan de sondage** Π.

On appelle **variable de J. CORNFIELD** attachée à une **unité de sondage** ω_m de Ω la **variable indicatrice** c_m tq $c_m = 1$ si $\omega_m \in A$ et $c_m = 0$ sinon.

Autrement dit :

$$(1) \qquad c_m = \mathbf{1}_A(\omega_m) = \begin{cases} 1 \text{ si } \omega_m \in A, \\ 0 \text{ si } \omega_m \notin A, \end{cases}$$

où $\mathbf{1}_A : \Omega \mapsto \{0, 1\}$ est la variable indicatrice associée à l'échantillon A ($\mathbf{1}_A$ est donc aléatoire parce que A est tiré dans Ω selon le plan Π).

(ii) Si $(\mathcal{Y}, \mathcal{C})$ est un **espace d'observation** et $\eta : \Omega \mapsto \mathcal{Y}$ la variable étudiée (**caractère** associé à chaque unité), on pose :

$$(2) \qquad \begin{aligned} Y &= (Y_1 ,..., Y_M), && \text{avec } Y_m = \eta(\omega_m) \ (\forall\ m), \\ y &= (y_1 ,..., y_N), && \text{avec } y_n = \eta(a_n) \ (\forall\ n). \end{aligned}$$

Par suite, si l'on pose $c = (c_1 ,..., c_M)$, le M-uple c est aléatoire puisqu'il dépend de A, qui est un échantillon (d'unités) aléatoire. La **loi de probabilité** de c détermine donc celle du caractère η observé sur l'échantillon A, ie détermine la loi de y.

(iii) Certaines **caractéristiques** « empiriques » (aléatoires) de A peuvent souvent être exprimées à partir de (1). Ainsi, lorsque $(\mathcal{Y}, \mathcal{C}) = (\mathbf{R}, \mathcal{B}_{\mathbf{R}})$ (caractère numérique scalaire), on peut exprimer (la sommation portant sur la population) :

(a) la **moyenne empirique** $\overline{y}_N$ selon :

$$(3) \qquad \overline{y}_N = N^{-1} . \Sigma_{n=1}^{N} y_n = (\Sigma_{m=1}^{M} c_m)^{-1} . \Sigma_{m=1}^{M} c_m Y_m ,$$

avec $\Sigma_{m=1}^{M} c_m = \Sigma_{m=1}^{M} \mathbf{1}_A(\omega_m) = N$;

(b) la **variance empirique** selon :

$$(4) \qquad s_N^2 = (N-1)^{-1} \Sigma_{n=1}^{N} (y_n - \overline{y}_N)^2 = (N-1)^{-1} \Sigma_{m=1}^{M} c_m (Y_m - \overline{y}_N)^2,$$

où $\overline{y}_N$ est défini en (3).

Dans le cas où A est un échantillon sans remise (ie $A \in \mathcal{T}_N$), on établit que :

$$e_M' \, c \; = \; N,$$

(5) $$E \, c_m \; = \; E \, c_m^2 \; = \; N / M \qquad (m = 1,..., M),$$

$$E \, c_l \, c_m \; = \; M^{-1} \, (M-1)^{-1} \, N \, (N-1) \qquad (\text{si } m \neq l).$$

On en déduit les formules classiques :

(6)

$$E \; \overline{y}_N \; = \; \overline{Y} \qquad$$ (**moyenne** théorique, ie moyenne sur la population),

$$V \; \overline{y}_N \; = \; N^{-1} \, (N-1)^{-1} \, (M-N) \, . \, \sigma^2,$$

où $\sigma^2 = S_M^2 = M^{-1} \, . \, \Sigma_{m=1}^M \, (Y_m - \overline{Y}_M)^2$ est la **variance** théorique (non corrigée).