

TEST DES SÉQUENCES (ou TEST DES SÉRIES) (E, I2, N)

(10 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Le **test des séquences**, ou **test des suites**, est un **test non qui paramétrique** intervient dans le cadre du problème à un échantillon ou dans celui à deux échantillons. Ce test est fondé sur la notion de **séquence**.

(i) Dans le cadre du **problème à un échantillon**, on considère un **modèle statistique** $(\Omega, \mathcal{F}, \mathcal{P})$ et une suite de **vars** $X = (X_1, \dots, X_N)$, où $X_n : \Omega \mapsto \mathbf{R}$ pour tout $n = 1, \dots, N$.

On suppose que X est une **suite indépendante** et l'on note $P^{X(n)}$ la loi de X_n et F_n sa **fonction de répartition** ($n = 1, \dots, N$) (où les $X(n)$ désignent les X_n).

On veut tester l'**hypothèse d'équidistribution** de X , ie l'**hypothèse de base** :

$$(1) \quad H_0 : P^{X(1)} = \dots = P^{X(N)},$$

(cf **suite équidistribuée**), contre une **alternative** de **tendance** tq eg :

$$(2) \quad H_1 : X \text{ est stochastiquement croissante,}$$

ie $\alpha < \beta \Rightarrow F_\alpha \geq F_\beta, \forall (\alpha, \beta) \in (N_N^*)^2$ (cf **croissance stochastique, test des rangs séquentiels**).

Soit $Z = (Z_1, \dots, Z_N)$ la suite des séquences de +1 et de -1 associée à X , et soit T_N le nombre total des séquences de la suite Z . On dit que T_N est la **statistique des séquences**, ou **statistique des séries**.

A « distance finie » (ie lorsque $N \ll +\infty$), on montre que T_N admet, sous l'hypothèse H_0 , les deux premiers **moments** suivants :

$$(3) \quad \begin{aligned} E_0 T_N &= (2N - 1) / 3, \\ V_0 T_N &= \{\sigma_0(T_N)\}^2 = (16N - 39) / 90, \end{aligned}$$

où $\sigma_0(S_N)$ désigne l'**écart-type** de S_N .

Lorsque la taille de l'échantillon est importante (ie $N \gg 0$), le test des rangs est défini à partir des **régions critiques** (de seuil α donné) qui se déduisent de la **propriété asymptotique** suivante (**convergence en loi**) :

$$(4) \quad \mathcal{L}(U_N) \xrightarrow{H_0}_{N \rightarrow +\infty} \mathcal{N}(0, 1) \quad (\text{loi normale réduite}),$$

où $U_N = (T_N - E_0 T_N) / \sigma_0(T_N)$ (statistique centrée réduite).

(ii) Dans le cadre du problème à deux échantillons (cf **problème à plusieurs échantillons**), on teste l'hypothèse de base (**homogénéité** stricte) :

$$(5) \quad H_0 : P_2^\xi = P_1^\xi$$

à l'aide de deux **échantillons aléatoires** indépendants, X^1 et X^2 (cf **indépendance, échantillon indépendant**), respectivement engendrés par les **vars** ξ_1 et ξ_2 .

L'**hypothèse alternative** est l'hypothèse « complémentaire » (ou « hypothèse omnibus ») $H_1 : P_2^\xi \neq P_1^\xi$.

On note $X = (X^1, X^2)$ l'**échantillon d'ensemble**, ou échantillon « empilé », ou échantillon « global », et l'on suppose X^2 indépendant de X^1 . On note $X^{(\cdot)}$ l'échantillon ordonné associé à X (cf **statistique d'ordre**) et $N = N_1 + N_2$ le nombre total d'**observations**.

On considère alors l'ensemble e des **indices** des coordonnées $X^{(n)}$ de $X^{(\cdot)}$ qui sont égales aux coordonnées $X_{1,n(1)}$ de $X^1 = (X_{1,1}, \dots, X_{1,N(1)})$, ie :

$$(6) \quad e = \{n \in N_N^* : \text{il existe } n_1 \in N_{N(1)}^* \text{ tq } X^{(n)} = X_{1,n(1)}\}.$$

On note E l'ensemble constitué de tous les ensembles e , avec $\text{Card } E = C_N^{N(1)}$ (nombre de **combinaisons** de N_1 éléments parmi N).

Sous l'hypothèse H_0 , E est doté de la **loi uniforme discrète** définie par $p(e) = (\text{card } E)^{-1}$, $\forall e \in E$.

Par ailleurs, on définit la **séquence des coordonnées** de X^1 dans X comme étant le plus grand nombre de coordonnées consécutives de $X^{(\cdot)}$ qui sont égales à des coordonnées de X^1 . On note $S_{N(1)N(2)}$ la **statistique** définie comme le **nombre de séquences** de coordonnées de X^1 .

On appelle alors **test des séquences**, ou **test des suites**, ou parfois **test des séries**, un test de l'hypothèse H_0 basé sur la statistique $S_{N(1)N(2)}$.

Une **région critique** de niveau $\alpha \in]0, 1[$ pour ce test est de la forme $w = [S_{N(1)N(2)} \leq q_{1-\alpha}]$, où $q_{1-\alpha}$ est le **quantile** d'ordre $1 - \alpha$ de la loi de $S_{N(1)N(2)}$.

Si $N_1 \leq N_2$ (ce qui ne restreint en rien la généralité), la **loi de probabilité** (à distance finie) P_0 de $S_{N(1)N(2)}$ est la suivante, toujours sous l'hypothèse H_0 :

(a) pour tout $n_1 \in \{1, \dots, N_1\}$:

$$(7) \quad P_0([S_{N(1)N(2)} = 2n_1]) = 2(A \cdot B) / D,$$

avec $A = C_{N(1)-1}^{n(1)-1}$, $B = C_{N(2)-1}^{n(1)-1}$, $D = C_{N(1)+N(2)}^{N(1)}$ (où C_r^s désigne le nombre de combinaisons de s éléments parmi r) ;

(b) pour tout $n_1 \in \{1, \dots, N_1 - 1\}$:

$$(8) \quad P_0 ([S_{N(1)N(2)} = 2 n_1 + 1]) = (C \cdot B + A \cdot E) / D,$$

avec $E = C_{N(2)-1}^{n(1)}$;

(c) pour tout $N_1 \leq N_2 - 1$:

$$(9) \quad P_0 ([S_{N(1)N(2)} = 2 N_1 + 1]) = C_{N(2)-1}^{N(1)} / D ;$$

(d) dans les autres cas, $P_0 ([S_{N(1)N(2)} = s]) = 0$.

Sous l'hypothèse H_0 , les deux premiers moments sont alors :

$$(10) \quad \begin{aligned} E_0 S_{N(1)N(2)} &= 1 + 2 (N_1 + N_2)^{-1} N_1 \cdot N_2 , \\ V_0 S_{N(1)N(2)} &= 2 (N_1 + N_2)^{-2} (N_1 + N_2 - 1)^{-1} N_1 \cdot N_2 (2 N_1 \cdot N_2 - N_1 - N_2). \end{aligned}$$

La tabulation de la loi permet, pour un seuil $\alpha \in]0, 1[$ donné, de réaliser le test à distance finie, ie $\max (N_1, N_2) << +\infty$.

Par suite (toujours sous H_0), on a la **propriété asymptotique** suivante (**convergence légale**) :

$$(11) \quad \mathcal{L}(U_{N(1)N(2)}) \rightarrow_{N^* \rightarrow \infty} \mathcal{N}(0, 1) \quad (\text{loi normale réduite}),$$

où $U_{N(1)N(2)} = (S_{N(1)N(2)} - E_0 S_{N(1)N(2)}) / \sigma_0 (S_{N(1)N(2)})$ désigne la statistique centrée et réduite (cf **normalisation**), $\sigma_0 (S_{N(1)N(2)}) = \{V_0 S_{N(1)N(2)}\}^{1/2}$ désigne l'**écart-type** de $S_{N(1)N(2)}$ et $N^* = \min (N_1, N_2)$.

Ceci permet de définir le test des séquences (asymptotique), lorsque les observations sont en nombre suffisant ($N_1 \gg 0$ et $N_2 \gg 0$).