TEST DU CHI-DEUX (C7, D1, I2)
(24 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

Il existe plusieurs **tests du chi-deux**, chacun relevant d'un contexte particulier, mais tous fondés sur la même **loi de probabilité**.

De façon générale, un **test du chi-deux** (ou $\mathcal{X}^2$) est un **test d'hypothèses** associé à une **statistique de test** dont la loi est la **loi du chi-deux**, que ce soit à distance finie ou à distance infinie (cf **loi asymptotique**).

L'expression se réfère, plus particulièrement à deux tests courants :

(a) un **test d'adéquation** : ce test vérifie la **concordance** de **données** avec une **loi de probabilité** ;

(b) un **test d'indépendance** : ce test vérifie l'**indépendance** interne à un **tableau statistique**, ie l'indépendance entre les **critères** qui définissent ce tableau (principalement, un **tableau de contingence**) (cf aussi **loi multivariée**, **test d'indépendance du chi-deux**).

(i) **Test d'adéquation d'une loi à des observations**. Ce test du chi-deux est basé sur la propriété suivante. Soit $X \sim \mathcal{M}_K (N, p)$ un **vecteur aléatoire** multinomial à partir duquel on définit une statistique, appelée **statistique de K. PEARSON** ou **statistique du chi-deux** :

(1) $K_N (p) = \Sigma_{k=1}^{K} (N p_k)^{-1} (X_k - N p_k)^2$, avec $e_K' X = N$.

Cette statistique vérifie la **propriété asymptotique** suivante (**convergence en loi**), qui explique l'origine de son nom :

(2) $\mathcal{L}(K_N (p)) \rightarrow_{N \rightarrow \infty} \mathcal{X}_{K-1}^2$ (**loi du chi-deux** à K-1 **degrés de liberté**).

On considère alors un **modèle image** $(\mathcal{X}, \mathcal{B}, (P_\theta^X)_{\theta \in \Theta})$ dans lequel $\mathcal{X}$ est l'ensemble puissance $\mathcal{X}_0^N = \mathbf{R}^N$ (avec $\mathcal{X}_0 = \mathbf{R}$) ainsi qu'une **partition** finie $\Pi_{\mathcal{X}} = (B_k)_{k=1,...,K}$ de $\mathcal{X}_0$, et l'on pose :

(3) $p = (p_1 ,..., p_K)$, avec $p_k = P_\theta ([X_k \in N_k])$, $\forall k \in N_K^*$.

Par suite, $p \in S_K$ (**simplexe** de $\mathbf{R}^K$) dépend de $\theta \in \Theta$, ce qui est aussi noté $p = p (\theta)$. Pour tester l'**hypothèse de base** d'**adéquation** :

(4) $H_0 : p = p_0$ (p_0 donné),

contre l'**hypothèse alternative** complémentaire :

(5) $H_1 : p \neq p_0$, (p_1 donné),

on utilise la **statistique de test** K_N (p) définie en (1) et l'on fonde le **test du chi-deux de K. PEARSON** sur la propriété (2). Pour un seuil critique $\alpha \in]0, 1[$ donné, ceci conduit à des **régions critiques** de la forme :

(6) $w = [K_N (p) \geq q_{1-\alpha}]$,

où $q_{1-\alpha}$ est le **quantile** d'ordre 1 - α de la loi $\mathcal{X}_{K-1}^2$.

(ii) En pratique, on ne connaît pas p = p (θ) puisqu'on ne connaît pas la « vraie » loi P_θ^X qui gouverne l'observation X. Lorsque $\Theta \subset \mathbf{R}^Q$, la « **règle de R.A. FISHER** » consiste à :

(a) estimer θ par un estimateur $\theta_N\tilde{}$ à valeurs dans Θ (eg l'**estimateur du maximum de vraisemblance**), ce qui conduit à l'**estimateur** p ($\theta_N\tilde{}$) ;

(b) calculer la statistique « corrigée » (ou estimée) (cf **correction**) :

(7) $K_N' (p) = \Sigma_{k=1}^K \{N\, p_k(\theta_N\tilde{})\}^{-1} \{X_k - N\, p_k(\theta_N\tilde{})\}^2$, avec $e_K' X = N$;

(c) effectuer le test de H_0 compte tenu de la propriété suivante (**théorème de C.H. CRAMER**), valable sous des conditions générales et lorsque l'**équation de vraisemblance** possède une solution unique :

(8) $\mathcal{L} \{K_N' (p(\theta_N\tilde{}))\} \rightarrow^{H0}{}_{N \to \infty} \mathcal{X}_{K-1}^2$ (**loi du chi-deux** à K-1 **degrés de liberté**).

Le test du chi-deux précédent est un **test « omnibus »** car l'hypothèse alternative H_1 est composite (ou « indéfinie »). Pour apprécier sa **puissance** (ie la probabilité de rejeter H_0 lorsque H_1 est vraie), il convient de spécifier certains éléments de H_1 . Par exemple, si l'on restreint le test du chi-deux aux **alternatives** H_1 tq :

(9) $H_1 : p = p_1$, avec $p_1 = p_0 + N^{-1/2} q$,

où q est tq $e_K' q = 0$, on montre que, si H_1 est vraie, la statistique $K_N' (p(\theta_N\tilde{}))$ définie en (7) suit une loi du chi-deux décentrée (cf **loi du chi-deux non centrale**), dont le paramètre de non **centralité** est $\lambda^2 = \Sigma_{k=1}^K (q_k^2 / p_{k,0}) = q' P_0^{-1} q$, avec $P_0 = \text{diag } p_0 = \text{diag } \{p_{1,0}, ..., p_{K,0}\} \in D_n (\mathbf{R})$ (**matrice diagonale**).

(iii) Le test d'adéquation précédent est asymptotiquement équivalent au **test du rapport des vraisemblances**, aussi appelé **test de l'entropie de C.E. SHANNON**, ou **test de l'information de C.E. SHANNON**, lequel est fondé sur la **statistique de test** suivante :

(17) $I_K (p) = \Sigma_{k=1}^K X_k . \text{Log} \{X_k / (N\, p_k)\} = \Sigma_{k=1}^K X_k . \text{Log} \{1 + (X_k - N\, p_k) / (N\, p_k)\}$.

Il en existe plusieurs version similaires ou dérivées (cf **statistique de HELLINGER**, **statistique de KULLBACK-LEIBLER**, **statistique du chi-deux modifiée**).

(iv) **Test d'indépendance** dans un **tableau statistique**. Le test du chi-deux consiste alors à tester l'indépendance entre des **variables aléatoires** qui sont des **caractères statistiques** (ou « **critères** ») définissant la **structure** du tableau (cf aussi **loi qualitative**). Ces variables peuvent être aussi bien des **variables quantitatives** que des **variables qualitatives**.

Dans le cas de deux critères, soit $T = (t_{ij})_{(i,j)}$ un tableau statistique d'ordre (m,n) tq $t_{ij} \geq 0$ pour tout $(i, j) \in \{1 ,..., m\} \times \{1,..., n\}$. On note p_{ij} la probabilité générant les **observations** de la « **case** » (ou « **cellule** ») (i, j) : la **suite** double $(p_{ij})_{i=1,...,m \; ; \; j = 1 ,..., n}$ est la **loi bivariée** associée au tableau (cf **loi multivariée**).

On teste l'hypothèse d'**indépendance** stochastique suivante :

(10) $H_0 : p_{ij} = p_{i.} \, p_{.j}$, $\quad \forall (i, j) \in N_m^* \times N_n^*$,

avec $p_{i.} = \Sigma_{j=1}^{n} p_{ij}$ $(i \in N_m^*)$, $p_{.j} = \Sigma_{i=1}^{m} p_{ij}$ $(j \in N_m^*)$ (**lois marginales** du tableau) et $p.. = \Sigma_{i=1}^{m} \Sigma_{j=1}^{n} p_{ij} = 1$ (total général).

Il existe donc m + n - 2 **paramètres** à estimer. En pratique, on estime resp $p_{i.}$ et $p_{.j}$ à l'aide les **rapports** empiriques :

(11)
$$p_{i.} = t_{i.} / t.. \,,$$
$$p_{.j} = t_{.j} / t.. \,,$$

avec des notations similaires pour T. La **statistique du chi-deux** qui fonde le test est alors la suivante :

(12) $K_{mn} = \Sigma_{i=1}^{m} \Sigma_{j=1}^{n} (a_{ij}^2 / b_{ij})$,

avec $a_{ij} = \{t_{ij} - t..^{-1} (t_{i.} \, t_{.j})\}$ et $b_{ij} = t..^{-1} (t_{i.} \, t_{.j})$.

Cette statistique est aussi appelée **carré moyen de contingence** lorsque $t_{ij} \in \mathbf{N}$ pour tout (i, j). Elle s'appelle aussi **g-statistique** et le test associé le **g-test**.

Sous des conditions générales, le test de H_0 se fonde sur la **propriété asymptotique** (**convergence légale**) suivante :

(13) $\mathcal{L}(K_{mn}) \rightarrow_{t^* \rightarrow +\infty} \mathcal{X}^2_{(m-1)(n-1)}$ (loi du chi-deux à (n-1)(k-1) **dl**),

où $t^* = \min_{(i,j)} t_{ij}$.

En particulier, si m = n = 2 (deux caractères à deux modalités chacun : « présence » ou « absence », « 1 » ou « 0 », etc), on note généralement le tableau T selon :

(14) $T = \{(a, b) / (c, d)\}$ (superposition de deux lignes : (a, b) et (c, d),

d'où la statistique de test :

(12)' $K_{22} = \{(a + b) (c + d) (a + c) (b + d)\}^{-1} (e_2' \, T \, e_2) (\text{dét } T)^2$,

où $e_2 = (1, 1)' \in \mathbf{R}^2$ (premier vecteur bissecteur). Par suite, $\mathcal{L}(K_{22}) = \mathcal{X}_1^2$.

(v) Ce qui précède se généralise directement à un **tableau statistique multidimensionnel** (eg k-dimensionnel) $T = (t_I)_{I \in \mathcal{T}}$, où $\mathcal{T}$ est l'ensemble produit des **indices**. La loi de probabilité sous-jacente est alors notée de façon conforme $P = (p_I)_{I \in \mathcal{T}}$.

La statistique de test utilisée est alors :

(15) $K_{n(1)...n(k)} = \Sigma_{I \in \mathcal{T}} (t_I - \tau_I)^2 / \tau_I$,

dans laquelle $\tau_I = \Pi_{\alpha=1}^{k} t.^{-1} \{\Sigma_{I(1) \in \mathcal{T}(1)} ... \Sigma_{I((\alpha-1) \in \mathcal{T}(\alpha-1)} \Sigma_{I((\alpha+1) \in \mathcal{T}(\alpha+1)} ... \Sigma_{I(k) \in \mathcal{T}(k)} t_I\}$, avec $t. = \Sigma_{I \in \mathcal{T}} t_I$ (total général) et où l'on note, par commodité, $I(\alpha)$ pour désigner l'indice I_α et $\mathcal{T}(\alpha)$ pour désigner l'ensemble $\mathcal{T}_\alpha$.

Sous l'hypothèse d'indépendance :

(16) $H_0 : p_I = \Pi_{\alpha=1}^{k} p_\alpha$,

on obtient :

(17) $\mathcal{L}(K_{n(1)...n(k)}) \rightarrow_{t^* \rightarrow +\infty} \mathcal{X}_d^2$,

avec $d = \Pi_{\alpha=1}^{k} (n_\alpha - 1)$.