Tests non paramétriques

Jean-Yves Tourneret[(1)]

(1) Université de Toulouse, ENSEEIHT-IRIT

jyt@n7.fr, http://perso.tesa.prd.fr/jyt/

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité

Bibliographie

- Notes de cours, polys, ...
  - C. Maugis-Rabusseau, Tests statistiques, la suite ..., Polycopié de cours INSA, 2019.
  - M. Fromont, Tests statistiques, rejeter, ne pas rejeter ... Se risquer ?, 2015-2016.
  - F.-G. Carpentier, Le test de Wilcoxon Mann Whitney, 2015.
  - R. Rakotomalala, Comparaison de populations. Tests non paramétriques, Université Lumière Lyon2, 2008.
- Livres
  - P. Capéraà and B. Van Cutsem, Méthodes et modèles en statistique, Dunod, Paris, 1988.
  - M. Hollander, D. A. Wolfe and E. Chicken, Nonparametric statistical methods, John Wiley & Sons, 2013.
  - W. J. Conover, Practical nonparametric statistics, John Wiley & Sons, 1999.
  - E. L. Lehmann and H. J. M. D'Abrera, Nonparametrics - Statistical methods based on ranks, Holden-day Inc., 1975.

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité

Généralités

- Principe : Un test statistique est un mécanisme qui permet de décider entre plusieurs hypothèses $H_0, H_1, ...$ à partir de n observations $x_1, ..., x_n$. On se limitera dans ce cours à deux hypothèses H_0 et H_1. Effectuer un test, c'est déterminer une statistique de test $T(X_1, ..., X_n)$ et un ensemble Δ tel que

$$\begin{aligned} &\mathcal{H}_0 \text{ rejetée si } T(X_1, ..., X_n) \in \Delta \\ &\mathcal{H}_0 \text{ acceptée si } T(X_1, ..., X_n) \notin \Delta. \end{aligned} \tag{1}$$

- Vocabulaire
  - H_0 est l'hypothèse nulle et H_1 est l'hypothèse alternative
  - $\mathcal{R} = \{(x_1, ..., x_n) | T(x_1, ..., x_n) \in \Delta\}$: région critique (ou région de rejet)
  - Fonction de test : $\phi(\boldsymbol{x}) = \mathbb{1}_{\mathcal{R}}(\boldsymbol{x})$, fonction indicatrice sur $\mathcal{R}$.

Définitions

- Tests paramétriques et non paramétriques
- Hypothèses simples et hypothèses composites
- Risque de première espèce = probabilité de fausse alarme

$$\alpha = \text{PFA} = P[\text{Rejeter } H_0 | H_0 \text{ vraie}]$$

- Risque de seconde espèce = probabilité de non-détection

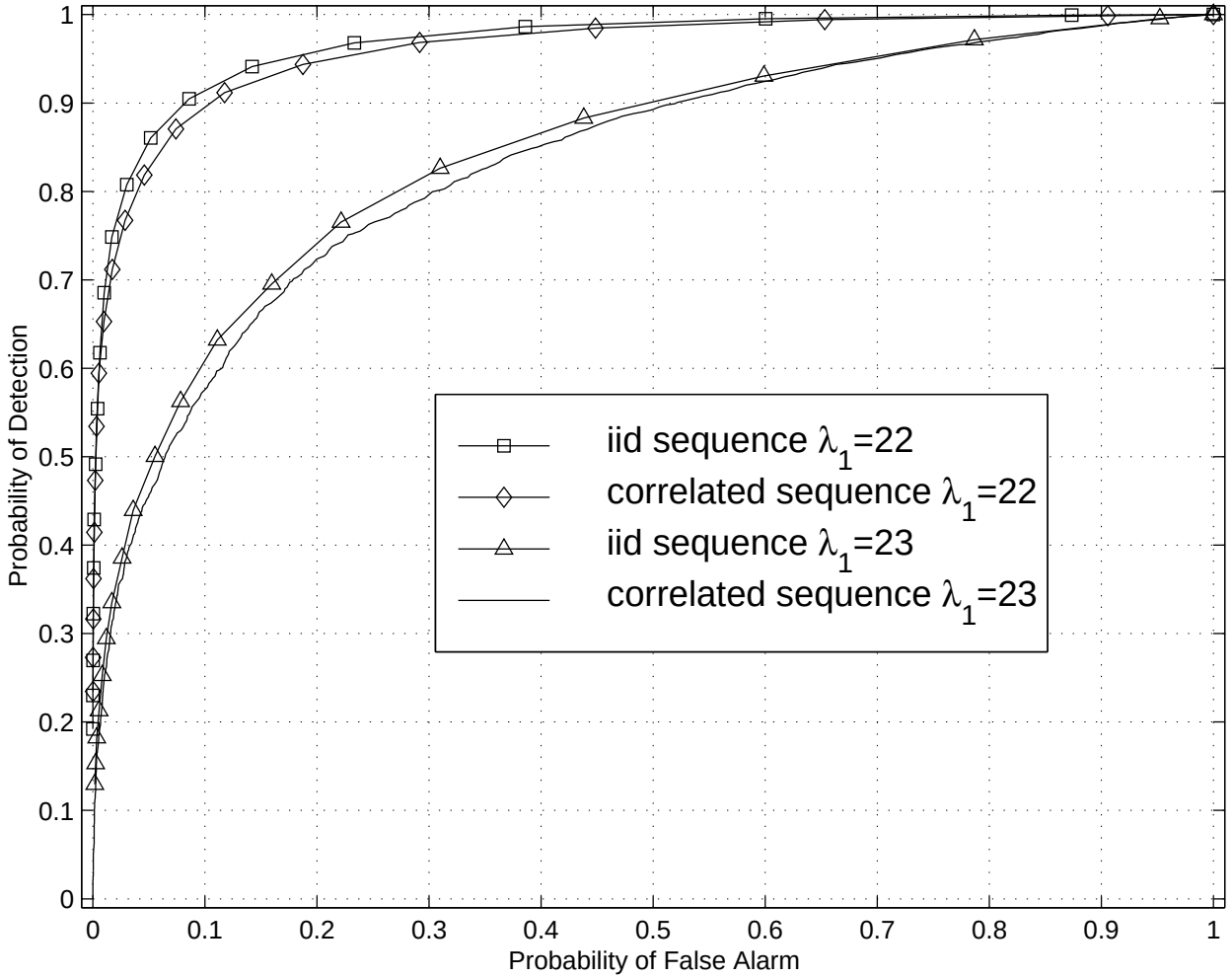
$$\beta = \text{PND} = P[\text{Rejeter } H_1 | H_1 \text{ vraie}]$$

- Puissance du test = probabilité de détection : $\pi = 1 - \beta$

Courbes COR

Caractéristiques opérationnelles du récepteur : PD vs PFA

ROC's for iid and correlated sequences



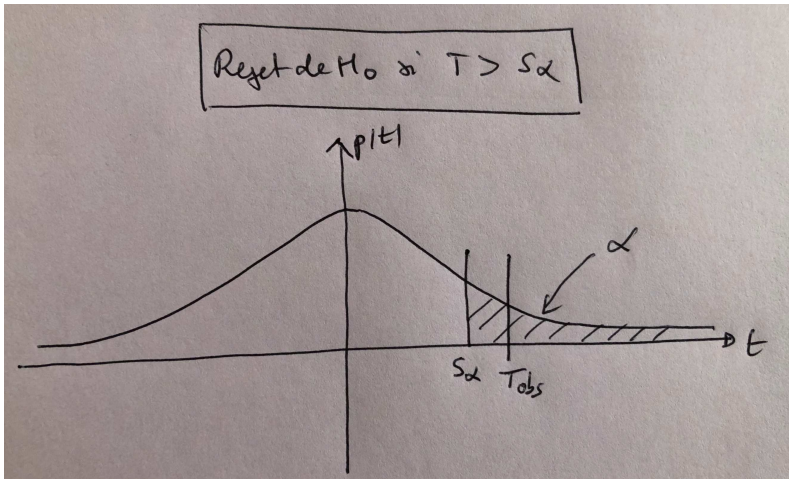
p-valeur d'un test

- Définition

$$p(\boldsymbol{x}) = \inf\{\alpha \in]0,1[\mid \boldsymbol{x} \in \mathcal{R}_\alpha\}$$

où $\mathcal{R}_\alpha$ est la zone de rejet pour α fixé et $\boldsymbol{x} = (x_1, ..., x_n)$. C'est la plus petite valeur de α pour laquelle on rejette H_0.

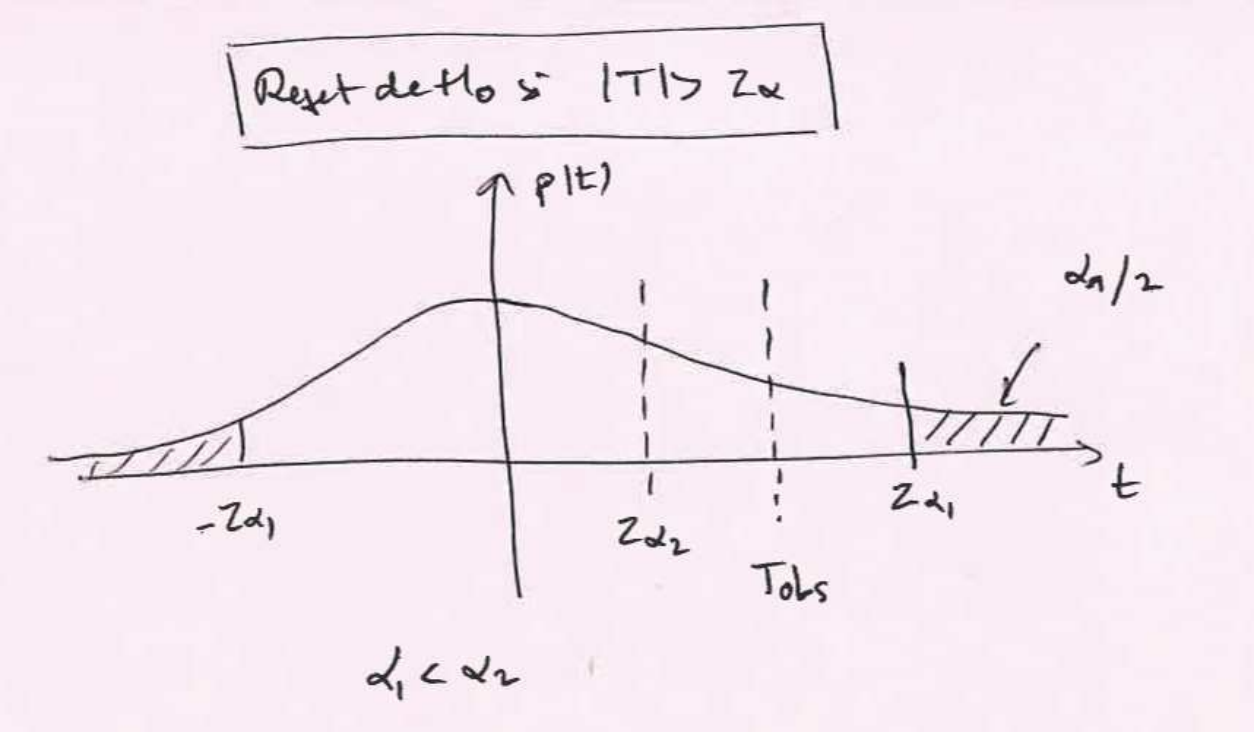
- Calcul



  - Si $\alpha = 0$, on accepte toujours H_0 donc $S_0 = +\infty$
  - Si $\alpha = 1$, on rejette toujours H_0 donc $S_1 = -\infty$
  - Plus petite valeur de α pour laquelle on rejette H_0

$$\alpha^* = 1 - F(T_{\text{obs}}).$$

Autre exemple



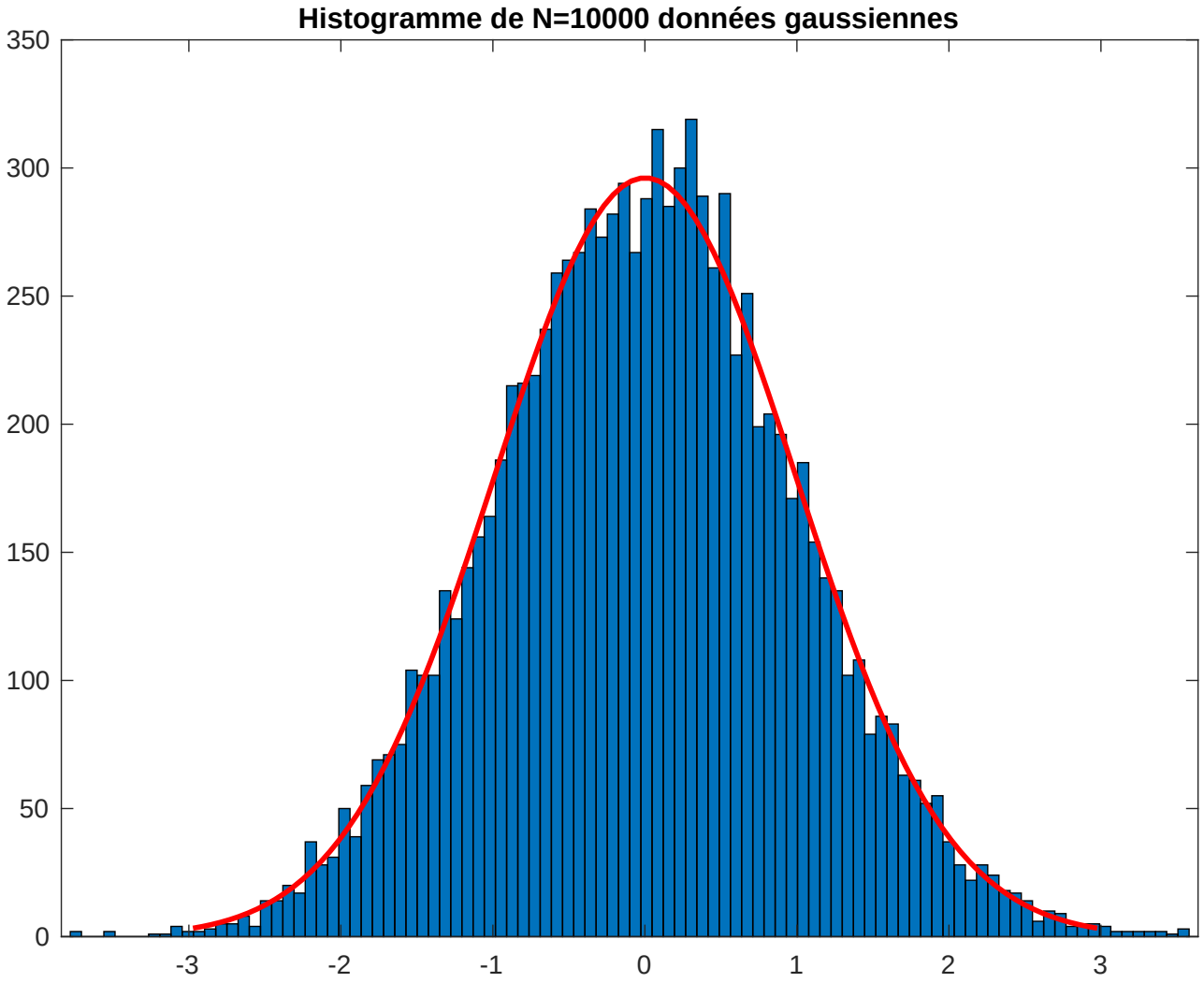
- Si $\alpha = 0$, on accepte toujours H_0 donc $z_0 = 0$
- Si $\alpha = 1$, on rejette toujours H_0 donc $z_1 = +\infty$
- Plus petite valeur de α pour laquelle on rejette H_0

$$\frac{\alpha^*}{2} = 1 - F(|T_{\text{obs}}|) \Leftrightarrow \alpha^* = 2[1 - F(|T_{\text{obs}})|].$$

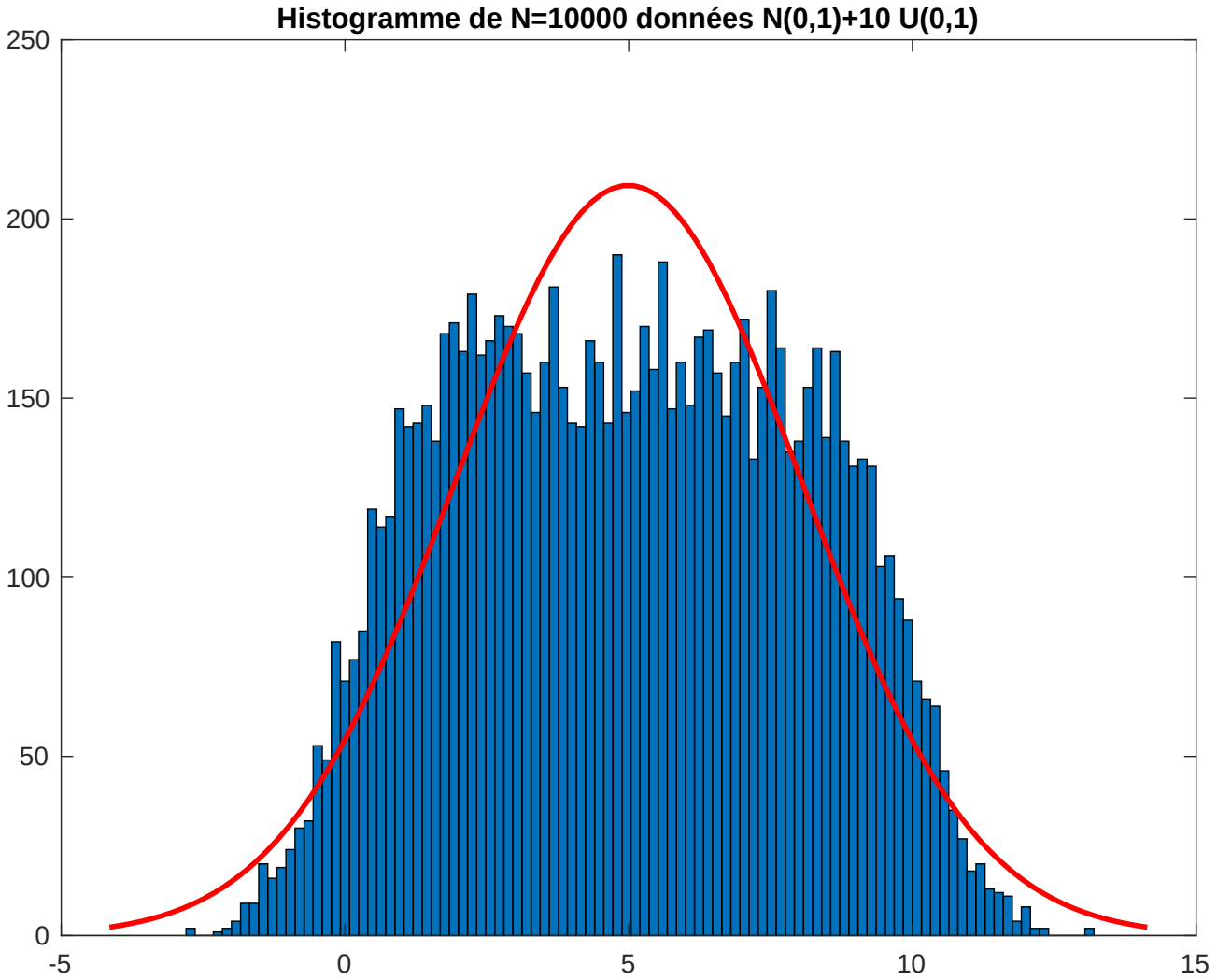
Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
  - Test de Kolmogorov
  - Test de Kolmogorov-Smirnov
  - Tests du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité

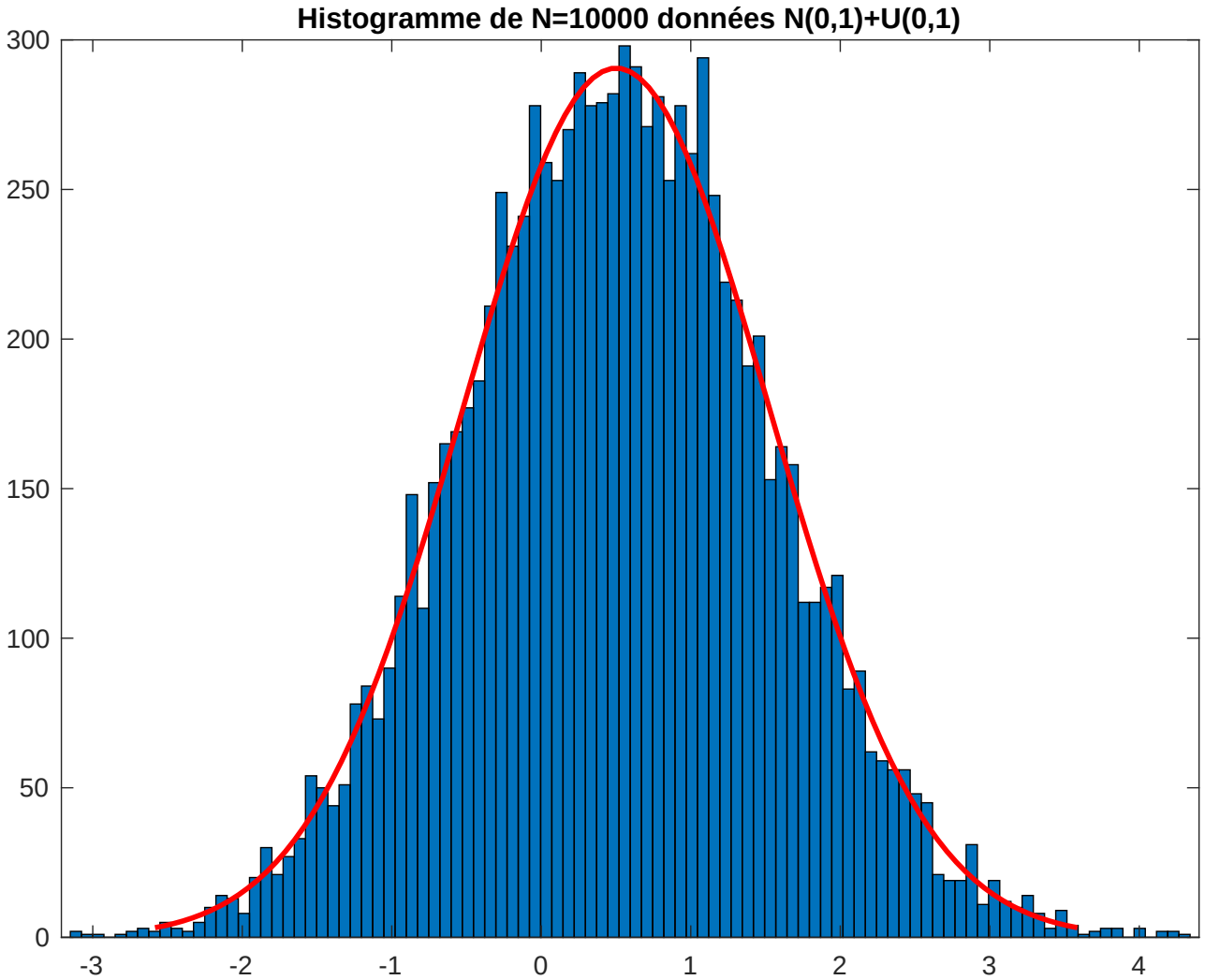
Motivations



Motivations



Motivations



Test de Kolmogorov

Le test de Kolmogorov est un test non paramétrique d'ajustement (ou d'adéquation) qui permet de tester les deux hypothèses suivantes

$$H_0 : L = L_0, \quad H_1 : L \neq L_0$$

où L_0 est une loi donnée. Le test consiste à déterminer si $(x_1, ..., x_n)$ est de loi L_0 ou non. On se limitera dans ce cours au cas simple où $x_i \in \mathbb{R}$.

- Définition

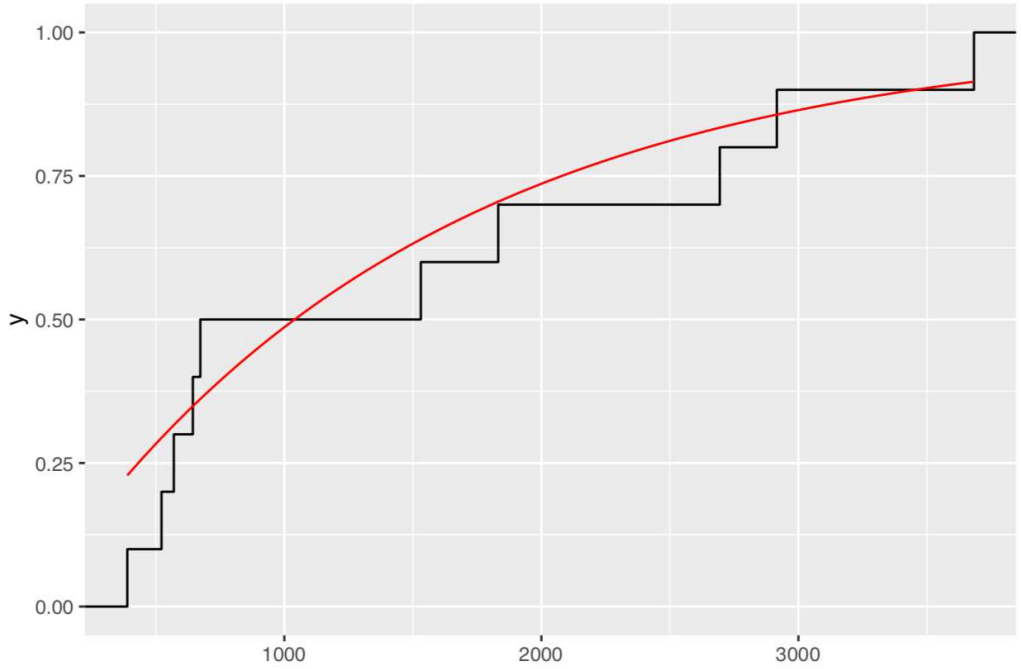
$$\text{Rejet de } H_0 \text{ si } D_n = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_0(x)| > S_{n,\alpha}$$

- Remarque : L_0 doit être une loi continue

Statistique de test

- Fonctions de répartition

  $F_0(x) = P[X \leq x]$ est la fonction de répartition théorique de L_0 et $\widehat{F}_n(x)$ est la fonction de répartition empirique de $(x_1, ..., x_n)$



D_n est l'écart maximum entre les deux courbes.

Calcul de D_n

- Á l'aide de l'échantillon ordonné

$$D_n = \max_{i \in \{1,\ldots,n\}} \max\{E_i^+, E_i^-\}$$

$$E_i^+ = \left|\widehat{F}_n\left(x_{(i)}^+\right) - F_0\left(x_{(i)}\right)\right|, \; E_i^- = \left|\widehat{F}_n\left(x_{(i)}^-\right) - F_0\left(x_{(i)}\right)\right|$$

- Rq : $x_{(1)}, ..., x_{(n)}$ est la statistique d'ordre de $x_1, ..., x_n$ telle que $x_{(1)} \leq ... \leq x_{(n)}$
- Rq : $\widehat{F}_n\left(x_{(i)}^+\right) = i/n$ et $\widehat{F}_n\left(x_{(i)}^-\right) = (i-1)/n$.

Statistique de test

- Loi de D_n sous H_0
  - Indépendante de L_0
  - Loi asymptotique

$$P[\sqrt{n}D_n < y] \underset{n\to\infty}{\to} 1-2\sum_{k=1}^{+\infty}(-1)^{k+1}\exp(-2k^2y^2) = K(y),\ y \geq 0$$

Convergence de cette série très rapide (pour $y > 0.56$, les trois premiers termes donnent une approximation avec une erreur inférieure à 10^{-4}).

- Détermination du seuil S_α

$$S_{n,\alpha} = \frac{1}{\sqrt{n}}K^{-1}(1-\alpha)$$

Le seuil dépend de α et de n.

Loi de D_n indépendante de L_0

- Propriétés de la fonction de répartition empirique

$$\begin{aligned}\widehat{F}_n(t) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i<t} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{F_0(X_i)<F_0(t)} \qquad (2) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i<F_0(t)} = F_{U,n}[F_0(t)],\end{aligned}$$

où $F_{U,n}(t)$ est la fonction de répartition empirique d'un échantillon de taille n de loi uniforme sur $[0,1]$ (si X_i suit la loi de fonction de répartition F_0).

Loi de D_n indépendante de L_0

- Loi de D_n

  Puisque $\widehat{F}_n(t) = F_{U,n}[F_0(t)]$, la statistique de test

  $$D_n = \sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F_0(t)|$$

  possède la même loi que

  $$\sup_{t \in \mathbb{R}} |F_{U,n}[F_0(t)] - F_0(t)| = \sup_{t \in [0,1]} |F_{U,n}(t) - t|$$

  sous l'hypothèse H_0. Cette loi est indépendante de la loi de l'échantillon $(X_1, ..., X_n)$!!

Remarques

- Puissance du test

  Non calculable

- Tests unilatéraux
  - Pour tester $H_0 : F = F_0$ contre $H_1 : F \geq F_0$, le test de Kolmogorov rejette H_0 si

$$D_n^+ = \sup_{t \in \mathbb{R}} \left[\widehat{F}_n(t) - F_0(t) \right] \geq S_{n,\alpha}$$

  - Pour tester $H_0 : F = F_0$ contre $H_1 : F \leq F_0$, le test de Kolmogorov rejette H_0 si

$$D_n^- = \sup_{t \in \mathbb{R}} \left[F_0(t) - \widehat{F}_n(t) \right] \geq S_{n,\alpha}$$

- Généralisation aux lois discrètes ou mixtes

  D. S. Dimitrova, V. K. Kaishev and S. Tan, “Computing the Kolmogorov-Smirnov distribution when the underlying cdf is purely discrete, mixed or continuous,” Journal of Statistical Software, vol. 95, no. 10, oct. 2020.

Exemple

Est-il raisonnable de penser que ces observations sont issues d'une population de loi uniforme sur $[0, 1]$?

x_i	0.0078	0.063	0.10	0.25	0.32	0.39	0.40	0.48	0.49	0.53
E_i^-	0.0078	0.013	0.00	0.10	0.07	0.14	0.05	0.008	0.04	0.03
E_i^+	0.0422	0.037	0.05	0.05	0.12	0.09	0.10	0.13	0.09	0.08
Max(E_i^+, E_i^-)	0.0422	0.037	0.05	0.1	0.12	0.14	0.10	0.13	0.09	0.08

x_i	0.67	0.68	0.69	0.73	0.79	0.80	0.87	0.88	0.90	0.996
E_i^-	0.17	0.13	0.04	0.03	0.04	0.05	0.07	0.03	0.05	0.046
E_i^+	0.12	0.08	0.09	0.08	0.09	0.00	0.02	0.02	0.00	$4e-3$
Max(E_i^+, E_i^-)	0.17	0.13	0.09	0.08	0.09	0.05	0.07	0.03	0.05	0.046

Exemple

- Statistique de test

$$D_n = 0.17$$

- Seuils pour $n = 20$

$S_{20,0.05}$	0.294
$S_{20,0.01}$	0.352

donc on accepte l'hypothèse H_0 avec les risques $\alpha = 0.01$ et $\alpha = 0.05$.

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
  - Test de Kolmogorov
  - Test de Kolmogorov-Smirnov
  - Test du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité

Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test non paramétrique qui permet de tester si deux échantillons $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ ont la même loi ou pas

$$H_0 : F = G, \quad H_1 : F \neq G$$

où F et G sont les fonctions de répartition respectives des deux échantillons et $\widehat{F}_n$ et $\widehat{G}_m$ leurs fonctions de répartition empiriques.

- Définition

$$\text{Rejet de } H_0 \text{ si } D_{n,m} = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - \widehat{G}_m(x)| > S_{n,m,\alpha}$$

- Propriété : si F est une loi continue, la loi de $D_{n,m}$ sous l'hypothèse nulle est indépendante de F (avec $F(t) = P[X_i \leq t]$).

Loi de $D_{n,m}$ sous H_0 indépendante de F

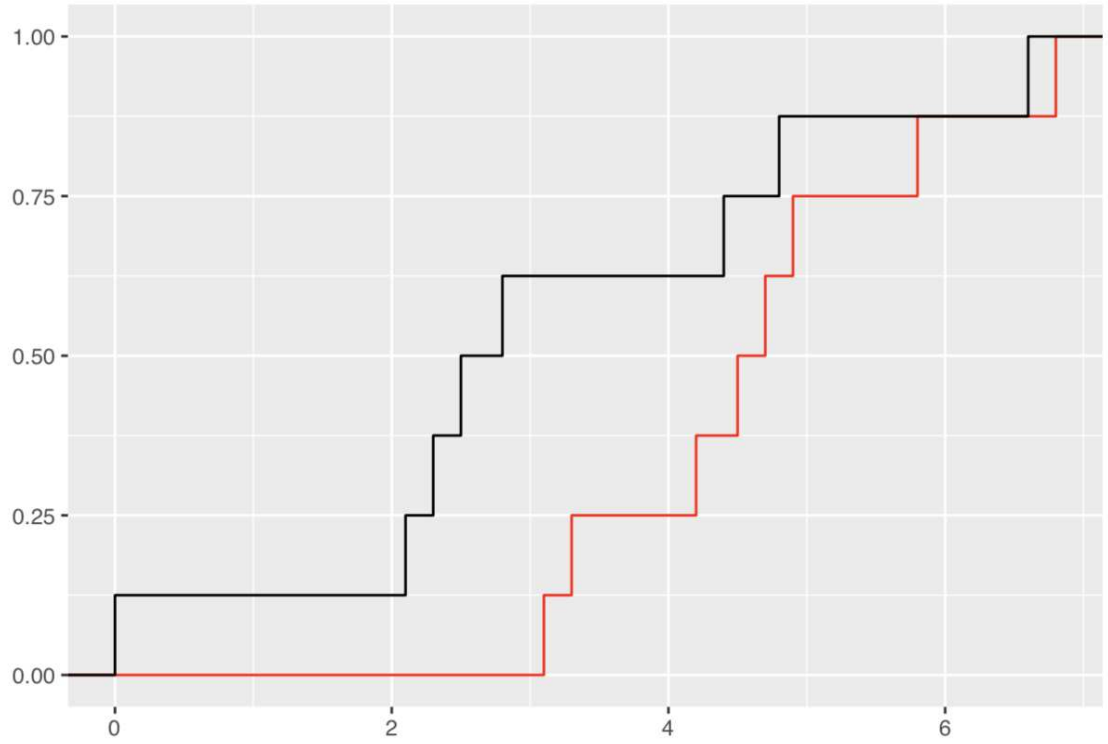
- Loi de $D_{n,m}$ sous l'hypothèse H_0

$$
\begin{aligned}
D_{n,m} &= \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)| \\
&= \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i\leq t} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{y_j\leq t} \right| \\
&= \sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{u_i\leq F(t)} - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{v_j\leq F(t)} \right| \\
&= \sup_{t\in[0,1]} |F_{U,n}(t) - F_{V,m}(t)|
\end{aligned}
\tag{3}
$$

qui est indépendante de la loi des échantillons $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$!!

Remarques

- Exemples de fonctions de répartition de x et y.



Calcul de la statistique de test

- Calcul de $D_{n,m}$

$$\begin{aligned} D_{n,m} &= \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)| \\ &= \max_{i\in\{1,\ldots,n+m\}} \left|\widehat{F}_n\left(z_{(i)}\right) - \hat{G}_m\left(z_{(i)}\right)\right| \end{aligned} \tag{4}$$

où $z_{(i)}$ est la $i$ème statistique d'ordre de la suite conjointe $(x_1, ..., x_n, y_1, ..., y_m)$.

Écritures équivalentes

- pour $D_{n,m}$

$$D_{n,m} = \max_{i \in \{1,\ldots,n+m\}} \frac{n+m}{nm} \left| \frac{im}{n+m} - \sum_{k=1}^{i} \alpha_k \right|$$

avec $\alpha_k = 1$ si la $k$ème plus petite observation appartient à la suite $\boldsymbol{y}$ et $\alpha_k = 0$ dans le cas contraire.

- pour $D_{n,n}$

$$D_{n,n} = \frac{1}{n} \max_{i \in \{1,\ldots,2n\}} \left| 2 \sum_{k=1}^{i} \alpha_k - i \right|$$

Preuve

On remarquera que $\sum_{k=1}^{n} \mathbb{1}_{X_k \leq z_{(i)}}$ est le nombre de x_k vérifiant $x_k \leq z_{(i)}$, i.e., le nombre de α_k égaux à 0 avant $z_{(i)}$, soit $\sum_{k=1}^{n} \mathbb{1}_{X_k \leq z_{(i)}} = \sum_{k=1}^{i}(1-\alpha_k)$. On en déduit

$$
\begin{aligned}
D_{n,m} &= \max_{i \in \{1,\ldots,n+m\}} \left| \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_k \leq z_{(i)}} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{Y_j \leq z_{(i)}} \right| \\
&= \max_{i \in \{1,\ldots,n+m\}} \left| \frac{1}{n} \sum_{k=1}^{i} (1-\alpha_k) - \frac{1}{m} \sum_{k=1}^{i} \alpha_k \right| \\
&= \max_{i \in \{1,\ldots,n+m\}} \left| \frac{i}{n} - \left(\frac{1}{n} + \frac{1}{m} \right) \sum_{k=1}^{i} \alpha_k \right| \\
&= \max_{i \in \{1,\ldots,n+m\}} \frac{n+m}{nm} \left| \frac{im}{n+m} - \sum_{k=1}^{i} \alpha_k \right|
\end{aligned}
\tag{5}
$$

Donc pour $n = m$, on a $\quad D_{n,n} = \dfrac{1}{n} \max\limits_{i \in \{1,\ldots,2n\}} \left| 2 \sum\limits_{k=1}^{i} \alpha_k - i \right|$

Exemple

Ces deux ensembles d'observations (les observations x_i sont en rouge, les y_j en bleu) sont-elles issues de la même loi ? (livre de Capéerà et Van Cutsem, page 111)

-4.53	-1.35	-0.96	-0.77	-0.42	-0.39	0.16	0.24	0.29	0.31
0.79	0.81	0.92	1.04	1.14	1.19	1.38	2.73	3.06	4.32

- Variables α_k

$$\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 0, \alpha_4 = 0, \alpha_5 = 0, \alpha_6 = 0, \alpha_7 = 1, \alpha_8 = 0, \alpha_9 = 0, \alpha_{10} = 0$$
$$\alpha_{11} = 1, \alpha_{12} = 1, \alpha_{13} = 1, \alpha_{14} = 1, \alpha_{15} = 0, \alpha_{16} = 0, \alpha_{17} = 0, \alpha_{18} = 1, \alpha_{19} = 1, \alpha_{20} = 1$$

- Valeurs de $\sum_{k=1}^{j} \alpha_k$: $1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 5, 6, 7, 7, 7, 7, 8, 9, 10$
- Valeurs de $\left|2\sum_{k=1}^{j} \alpha_k - j\right|$: $1, 2, 1, 0, 1, 2, 1, 2, 3, 4, 3, 2, 1, 0, 1, 2, 3, 2, 1, 0$
- Statistique de test : $D_{n,n} = 0.4$
- Seuil : $S_{10,0.05} = 0.5$, donc on accepte H_0 avec $\alpha = 0.05$.
- p-value : $p = 0.3129$. Donc on accepte H_0 avec des risques $\alpha \leq 0.3129$.

Loi de la statistique $D_{n,m}$

- Loi de $D_{n,m}$ pour $m = n$

$$P[D_{n,n} > d] = 2 \sum_{j=1}^{|\text{ent}(\frac{k}{n})|} (-1)^{j+1} \frac{(n!)^2}{(n-jk)!(n+jk)!}$$

où ent désigne la partie entière.

- Loi de $D_{n,m}$ pour $n \neq m$
  - Plus compliquée que celle de $D_{n,n}$ mais elle est tabulée
  - Loi asymptotique (convergence lente)

$$P\left[\sqrt{\frac{mn}{m+n}} D_{n,m} < y\right] \underset{n,m\to\infty}{\to} \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2) = K(y).$$

Remarques

- Test unilatéral

  Pour tester $F \geq G$, on peut rejeter H_0 si

$$D^+_{n,m} = \sup_{t \in \mathbb{R}} \left[\hat{F}_n(t) - \hat{G}_m(t) \right] > S_{n,m,\alpha}.$$

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
  - Test de Kolmogorov
  - Test de Kolmogorov-Smirnov
  - Tests du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité

Test d'ajustement du χ^2

Le test du χ^2 est un test non paramétrique d'ajustement (ou d'adéquation) qui permet de tester les deux hypothèses suivantes

$$H_0 : L = L_0, \quad H_1 : L \neq L_0$$

où L_0 est une loi donnée. Le test consiste à déterminer si $(x_1, ..., x_n)$ est de loi L_0 ou non. On se limitera dans ce cours au cas simple où $x_i \in \mathbb{R}$.

- Définition

$$\text{Rejet de } H_0 \text{ si } \phi_n = \sum_{k=1}^{K} \frac{(Z_k - np_k)^2}{np_k} > S_{K,\alpha}$$

- Remarque : L_0 peut être une loi discrète ou continue

Test d'ajustement du χ^2

- Statistique de test
  - Z_k : nombre d'observations x_i appartenant à la classe C_k, $k = 1, ..., K$
  - p_k : probabilité qu'une observation x_i appartienne à la classe C_k sachant $X_i \sim L_0$

$$P\left[X_i \in C_k | X_i \sim L_0\right]$$

  - n : nombre total d'observations
- Loi (asymptotique) de la statistique de test sous H_0

$$\phi_n \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \chi^2_{K-1}$$

Remarques

- Interprétation de ϕ_n

$$\phi_n = \sum_{k=1}^{K} \frac{n}{p_k} \left(\frac{Z_k}{n} - p_k \right)^2$$

  Distance entre probabilités théoriques et empiriques

- Loi asymptotique de ϕ_n : voir notes de cours ou livres
- Nombre d'observations fini
  Une heuristique dit que la loi asymptotique de ϕ_n est une bonne approximation pour n fini si 80% des classes vérifient $np_k \geq 5$ et si $p_k > 0, \forall k = 1, ..., K$
  ☞ Classes équiprobables

Loi asymptotique de ϕ_n

- Loi de $(Z_1, ..., Z_K)$

  Comme $Z_k = \sum_{i=1}^n \mathbb{I}_{C_k}(X_i)$, $(Z_1, ..., Z_K)^T$ suit sous l'hypothèse H_0 une loi multinomiale de moyenne $\boldsymbol{m} = (np_1, ..., np_K)^T$ et de matrice de covariance $\boldsymbol{\Sigma}$ avec

$$\begin{aligned}\boldsymbol{\Sigma}_{kl} &= E[Z_k Z_l] - E[Z_k]E[Z_l] \\ &= \sum_{i=1}^n \sum_{j=1}^n E[\mathbb{I}_{C_k}(X_i)\mathbb{I}_{C_l}(X_j)] - n^2 p_k p_l \qquad (6) \\ &= (n^2 - n)p_k p_l - n^2 p_k p_l = -np_k p_l\end{aligned}$$

  et

$$\boldsymbol{\Sigma}_{kk} = \text{var} Z_k = np_k(1 - p_k)$$

- Propriétés de $\left(\frac{Z_1 - np_1}{\sqrt{np_1}}, ..., \frac{Z_K - np_K}{\sqrt{np_K}}\right)$
  - moyennes nulles
  - variances $1 - p_k$, avec $k = 1, ..., K$
  - covariances $-\sqrt{p_k p_l}$, avec $k, l = 1, ..., K$ et $k \neq l$
  - somme de n vecteurs indépendants et de même loi

Loi asymptotique de ϕ_n

Donc d'après le théorème central limite multivarié

$$\left(\frac{Z_1 - np_1}{\sqrt{np_1}}, ..., \frac{Z_K - np_K}{\sqrt{np_K}} \right) \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

avec $\mathbf{\Sigma} = \boldsymbol{I}_K - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^T$.
Mais $\boldsymbol{I}_K - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^T$ est la matrice de projection orthogonale sur vec$\{\sqrt{\boldsymbol{p}}\}^\perp$ qui est de dimension $K - 1$. Donc, d'après le théorème de Cochran avec $\boldsymbol{m} = \mathbf{0}$ et $\sigma^2 = 1$, on a

$$\sum_{k=1}^{K} \frac{(Z_k - np_k)^2}{np_k} \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} \chi^2_{K-1}$$

Rappels

- Théorème central limite multivarié

  Si $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ est une suite de vecteurs aléatoires de $\mathbb{R}^p$ indépendants et de même loi de vecteur moyenne $\boldsymbol{m} \in \mathbb{R}^p$ et de matrice de covariance $\boldsymbol{\Sigma} \in \mathcal{M}_p(\mathbb{R})$, alors

$$\frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} X_i - n\boldsymbol{m} \right) \underset{n \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- Théorème de Cochran
  - Hypothèses

    Soit $\boldsymbol{X}$ un vecteur Gaussien de loi $\mathcal{N}_n(\boldsymbol{m}, \sigma^2 \boldsymbol{I}_n)$, où $\boldsymbol{m} \in \mathbb{R}^n$, $\sigma^2 > 0$ et $\boldsymbol{I}_n$ la matrice identité de taille $n \times n$. Soient p sous-espaces vectoriels orthogonaux $E_1, ..., E_p$ de dimensions $d_1, ..., d_p$ tels que $\mathbb{R}^n = E_1 \oplus ... \oplus E_p$ et $\boldsymbol{Y}_k = \boldsymbol{P}_k \boldsymbol{X}$ la projection de $\boldsymbol{X}$ sur E_k ($\boldsymbol{P}_k$ matrice de projection orthogonale sur E_k).
  - Conclusions
    - Les vecteurs $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_p$ sont indépendants et $\boldsymbol{Y}_k \sim \mathcal{N}_n(\boldsymbol{P}_k \boldsymbol{m}, \sigma^2 \boldsymbol{P}_k)$
    - Les variables aléatoires $Z_k = \|\boldsymbol{Y}_k - \boldsymbol{P}_k \boldsymbol{m}\|^2$ sont indépendantes et $\frac{Z_k}{\sigma^2} \sim \chi^2_{d_k}$.

Remarques

- Correction
  Lorsque les paramètres de la loi L_0 sont inconnus

$$\phi_n \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \chi^2_{K-1-n_p}$$

  où n_p est le nombre de paramètres inconnus estimés par la méthode du maximum de vraisemblance

- Constitution des classes dans le cas d'une loi discrète
- Puissance du test
  Non calculable

Exemple

4.13	1.41	−1.16	−0.75	1.96	2.46	0.197	0.24	0.42	2.00
2.08	1.48	1.73	0.82	0.33	−0.76	0.42	4.60	−2.83	0.197
2.59	0.54	4.06	−0.69	4.99	0.67	2.45	5.61	2.13	1.76
5.03	0.85	1.29	0.17	−0.38	2.76	−1.03	1.87	4.48	0.73

Est-il raisonnable de penser que ces observations sont issues d'une population de loi $\mathcal{N}(1, 4)$?

Solution

- Classes

$$C_1 :]-\infty, -0.34], C_2 :]-0.34, 1], C_3 :]1, 2.34], C_4 :]2.34, \infty[$$

- Nombres d'observations

$$Z_1 = 7,\ Z_2 = 12,\ Z_3 = 10,\ Z_4 = 11$$

Exemple

- Statistique de test

$$\phi_n = 1.4$$

- Seuils

	χ^2_2	χ^2_3
$S_{0.05}$	5.991	7.815
$S_{0.01}$	9.210	11.345

donc on accepte l'hypothèse H_0 avec les risques $\alpha = 0.01$ et $\alpha = 0.05$.

Test du χ^2 d'indépendance

- Principe : le test du χ^2 d'indépendance permet de tester si deux variables aléatoires réelles X et Y admettant un nombre fini de modalités $\{a_1, ..., a_k\}$ et $\{b_1, ..., b_L\}$ sont indépendantes ou pas, à partir d'un échantillon $\{(\boldsymbol{X}_1, \boldsymbol{Y}_1), ..., (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$ de ces variables aléatoires. Les deux hypothèses sont

$$H_0 : X \text{ et } Y \text{ indépendantes}, \quad H_1 : X \text{ et } Y \text{ non indépendantes}$$

Test du χ^2 d'indépendance

- Stratégie de test

  Rejet de H_0 si

$$I_n = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.} N_{.,l}}{n} \right)^2}{\frac{N_{k,.} N_{.,l}}{n}} > S_{K,L,\alpha}$$

  avec

  - $N_{k,l}$ est le nombre de (x_i, y_j) tels que $x_i \in C_k = \{a_k\}$ et $y_j \in \Omega_l = \{b_l\}$
  - $N_{k,.}$ est le nombre de x_i tels que $x_i \in C_k$
  - $N_{.,l}$ est le nombre de y_j tels que $y_j \in \Omega_l$ où l'ensemble des valeurs de X est divisé en K classes $C_1, ..., C_K$ et l'ensemble des valeurs de Y est divisé en L classes $\Omega_1, ..., \Omega_L$.
  - K : nombre de valeurs possibles de la première variable X
  - L : nombre de valeurs possibles de la seconde variable Y

Test du χ^2 d'indépendance

- Justification

  Sous H_0, on a $P[X \in C_k, Y \in \Omega_l] = P[X \in C_k] \times P[Y \in \Omega_l]$, soit $\frac{N_{k,l}}{n} \approx \frac{N_{k,.}}{n} \frac{N_{.,l}}{n}$, ce qui induit que la distance suivante est petite

$$n \frac{\left(\frac{N_{k,l}}{n} - \frac{N_{k,.}}{n} \frac{N_{.,l}}{n}\right)^2}{\frac{N_{k,.}}{n} \frac{N_{.,l}}{n}} = \frac{\left(N_{k,l} - \frac{N_{k,.} N_{.,l}}{n}\right)^2}{\frac{N_{k,.} N_{.,l}}{n}}$$

- Loi de la statistique de test

  La loi asymptotique de I_n sous l'hypothèse H_0 est une loi du χ^2 à $(K-1)(L-1)$ degrés de liberté (ddl). Le nombre de ddl est $(KL-1) - [(K-1)+(L-1)] = (K-1)(L-1)$. Sous H_0, on a KL probabilités vérifiant $P[X_i \in C_k, Y_j \in \Omega_l] = P[X_i \in C_k] P[Y_j \in \Omega_l]$, K probabilités inconnues $P[X_i \in C_k]$ et L probabilités inconnues $P[Y_j \in \Omega_l]$.

- Remarques
  - Lorsque les lois de X et Y sont discrètes avec un nombre infini de valeurs ou continues, on construit les classes C_k et Ω_l en regroupant plusieurs valeurs discrètes ou à l'aide d'intervalles.
  - Les effectifs de chaque classe doivent être ≥ 5 et le nombre total d'observations doit vérifier $n \geq 30$

Exemple

Une enquête a été réalisée auprès d'un échantillon de 250 personnes au sujet de l'abaissement à 16 ans du droit de vote. Les réponses ont été classées suivant le niveau d'instruction des personnes interrogées dans le tableau de contingence ci-dessous. Peut-on affirmer, au risque d'erreur de 5%, qu'il existe une relation entre l'opinion d'une personne sur cette question et son niveau d'instruction ?

Tableau de contingences

	Pour	Contre	$N_{k.}$
Brevet	10	15	25
Bac	20	85	105
Bac +2 et plus	20	100	120
$N_{.l}$	50	200	250

Exemple

Solution

- Statistique de test

$$I_n = \frac{\left(10 - \frac{25\times 50}{250}\right)^2}{\frac{25\times 50}{250}} + \ldots + \frac{\left(100 - \frac{120\times 200}{250}\right)^2}{\frac{120\times 200}{250}} = 7.14$$

- Seuil : $S_\alpha = F_2^{-1}(0.95) = 5.99$ car $(K-1)\times(L-1) = 2$
- Décision

  On rejette H_0 avec le risque $\alpha = 0.05$ (la p-valeur vaut 0.028) et donc on décide qu'il y a une relation entre opinion et niveau d'instruction

- Matlab

  [p,Q] = chi2test([10,15; 20,85; 20,100])

  Q=7.1429, pval=0.0281

Test du χ^2 d'homogénéité

- Principe : le test du χ^2 d'homogénéité permet de tester si L échantillons $(\boldsymbol{X}_{l,1}, ..., \boldsymbol{X}_{l,n_l})$ de lois discrètes définies sur le même support $\{a_1, ..., a_K\}$ ont la même loi ou pas. Si on note $\boldsymbol{\pi}_l = (\pi_{l,1}, ..., \pi_{l,K}) = (P[X_l = a_1], ..., P[X_l = a_K])$ la loi du $l$ème échantillon, les hypothèses sont

$$H_0 : \boldsymbol{\pi}_l = ... = \boldsymbol{\pi}_L (\text{les échantillons ont la même loi}), \quad H_1 : \text{non } H_0$$

Test du χ^2 d'homogénéité

- Stratégie de test

  Rejet de H_0 si

$$J_n = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.} N_{.,l}}{n} \right)^2}{\frac{N_{k,.} N_{.,l}}{n}} > S_{K,L,\alpha}$$

  avec

  - $N_{k,l} = \sum_{i=1}^{n_l} \mathbb{1}_{X_{l,i}=a_k}$, nombre de $x_{l,i}$ (de l'échantillon l) égaux à a_k
  - $N_{k,.} = \sum_{l=1}^{L} N_{k,l}$ est le nombre d'observations (des l échantillons) égales à a_k
  - $N_{.,l} = \sum_{k=1}^{K} N_{k,l} = n_l$: taille de l'échantillon l
  - $n = \sum_{l=1}^{L} n_l$: nombre total d'observations
  - K : nombre de valeurs prises par les échantillons
  - L : nombre d'échantillons

Test du χ^2 d'homogénéité

- Motivation

  Sous H_0, les l échantillons ont la même loi, donc la probabilité théorique d'avoir a_k est $\frac{N_{k,.}}{n}$. Donc l'effectif théorique de l'échantillon l pour a_k est $\frac{N_{k,.}}{n} \times N_{.,l}$ qui est comparé à l'effectif observé via

$$\frac{\left(N_{k,l} - \frac{N_{k,.} \times N_{.,l}}{n}\right)^2}{\frac{N_{k,.} \times N_{.,l}}{n}}$$

- Loi de la statistique de test

  La loi asymptotique de J_n sous l'hypothèse H_0 est une loi du χ^2 à $(K-1)(L-1)$ degrés de liberté.

- Nombre de données

  Les effectifs théoriques de chaque classe doivent être ≥ 5 et on doit avoir $n \geq 30$.

- Lois continues

  Dans le cas de deux lois continues, on peut découper le support des lois en intervalles et compter les nombres d'observations appartenant à ces intervalles.

Exemple

Dans cet exemple, on souhaite savoir si le taux de participation à un club sportif des élèves de deux collèges A et B est identique ou pas. On a donc deux échantillons $E_1 = (X_{1,1}, ..., X_{1,n_1})$ et $E_2 = (X_{2,1}, ..., X_{2,n_2})$ avec $X_{l,i} =$ participation du ième élève du collège $l \in \{a_1, a_2\} = \{$"oui", "non"$\}$ et on veut tester si les deux populations sont homogènes (hypothèse H_0) ou pas. Les effectifs observés sont les suivants

	Collège A	Collège B	$N_{k.}$
Oui	12	26	38
Non	38	34	72
$N_{.l}$	50	60	$n = 110$

Les effectifs théoriques sous l'hypothèse H_0 sont

	Collège A	Collège B
Oui	$\frac{38 \times 50}{110}$	$\frac{38 \times 60}{110}$
Non	$\frac{72 \times 50}{110}$	$\frac{72 \times 60}{110}$

Solution

Solution

- Statistique de test

$$J_n = \frac{\left(12 - \frac{38\times 50}{110}\right)^2}{\frac{38\times 50}{110}} + \ldots + \frac{\left(34 - \frac{72\times 60}{110}\right)^2}{\frac{72\times 60}{110}} = 4.504$$

- seuil : $S_\alpha = F_1^{-1}(0.95) = 3.84$ où F_1 est la fonction de répartition d'une loi du χ^2_1 car $(K-1)\times(L-1) = 1$
- Décision

  On rejette H_0 avec le risque $\alpha = 0.05$ et donc on décide que le taux de participation à un club sportif est différent entre les deux collèges (la p-valeur vaut $p = 0.034$).

Que faut-il savoir ?

- Principe et mise en oeuvre d'un test de Kolmogorov
- Principe et mise en oeuvre d'un test de Kolmogorov-Smirnov
- Principe et mise en oeuvre d'un test du χ^2 pour tester
  - l'adéquation d'un échantillon à une loi
  - l'indépendance de deux échantillons
  - l'homogénéité de plusieurs échantillons

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
- Chapitre 3 : Tests basés sur les rangs
  - Test de Wilcoxon-Mann Whitney
  - Test de la médiane
- Chapitre 4 : Tests de normalité

Test de Mann-Whitney

Le test de Mann-Whitney est un test non paramétrique qui permet de tester si deux échantillons indépendants $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ de fonctions de répartition F et G ont la même loi ou pas (avec $n \leq m$ par convention). Dans un premier temps, on considère les deux hypothèses

$$H_0 : F = G, \quad H_1 : G < F$$

- Définition

$$\text{Rejet de } H_0 \text{ si } U_y = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{Y_j > X_i} \geq S_{n,m,\alpha}$$

Remarques

- Autres tests
  - Pour tester $H_0 : F = G$ contre $H_1 : F \leq G$, le test de Mann-Whitney rejette H_0 si $U_y \leq S_{2,n,m,\alpha} = \frac{nm}{2} - S_{1,n,m,\alpha}$ (voir, e.g., [Hollander, p. 117]).
  - Pour tester $H_0 : F = G$ contre $H_1 : F \neq G$, le test de Mann-Whitney rejette H_0 si $U_y \geq S_{1,n,m,\frac{\alpha}{2}}$ ou $U_y \leq S_{2,n,m,\frac{\alpha}{2}} = \frac{nm}{2} - S_{1,n,m,\frac{\alpha}{2}}$ [Hollander, p. 117].
- Prise en compte des égalités

  La statistique du test de Mann-Whitney a été déterminée pour des lois continues. S'il y a des valeurs égales dans les deux séries, on utilise le test

$$\text{Rejet de } H_0 \text{ si } U_y = \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ \mathbb{1}_{Y_j > X_i} + \frac{1}{2} \mathbb{1}_{Y_j = X_i} \right\} \geq S_{n,m,\alpha}$$

Test U de Mann Whitney

La statistique de test U_y est basée sur les valeurs du second échantillon dépassant celles du premier échantillon. Si on échange les rôles des deux échantillons, on obtient une statistique de test notée U_x qui vérifie $U_x + U_y = nm$. On peut donc aussi utiliser la statistique de test U_x qui a la même loi que U_y sous H_0. En pratique, pour un test bilatéral, on utilise parfois le minimum entre U_x et U_y appelé test U de Mann-Whitney.

- Définition du test (bilatéral) U de Mann-Whitney
  Le test U de Mann-Whitney est basé sur la statistique de test $U = \min(U_x, U_y)$ et rejette l'hypothèse H_0 si $U < S_1$. Attention, les tables sont parfois associées à cette statistique $U = \min(U_x, U_y)$ et pas à U_y (ou à U_x) ![a].
- Propriétés
  - De manière évidente, $U \in \left[0, \frac{mn}{2}\right]$ ($U_x \in [0, mn]$ and $U_y = nm - U_x$).
  - On montre que la loi de U sous H_0 est liée à celle de U_y (ou U_x) grâce à la relation $P[U < u] = P[U_x < u \text{ ou } U_y = nm - U_x < u] = 2P[U_x < u]$ pour $u \in \left[0, \frac{mn}{2}\right]$.
  - D'après le point précédent, la loi de $U = \min(U_x, U_y)$ diffère de celle de U_y et les lois asymptotiques également.

[a] Sous Matlab, *ranksum(X,Y)* calcule la somme des ranks de X avec $n \leq m$.

Remarques

- Lien entre Wilcoxon (1945) et Mann-Whitney (1947)
  On montre que
  $$U_y = W_y - \frac{m(m+1)}{2}$$
  où $W_y = \sum_{j=1}^{m} S_j$ et S_j est le rang de Y_j parmi les $n+m$ données réunies $(X_1, ..., X_n, Y_1, ..., Y_m)$ (le minimum a le rang 1, la suivante le rang 2, ...) $\Rightarrow$ Test de Wilcoxon-Mann-Whitney.
  Remarque
  $$\frac{m(m+1)}{2} \leq W_y \leq (n+1) + (n+2) + ... + (n+m) = mn + \frac{m(m+1)}{2}.$$

- Correction de continuité
  Avec Matlab (ranksum.m) et R (wilcox.test), la correction de continuité est utilisée par défaut. Ce n'est pas le cas avec Python (stats.ranksums).

Lien entre Wilcoxon et Mann-Whitney

- Rang de Y_j

$$S_j = \sum_{k=1}^{n+m} \mathbb{1}_{Y_j \geq Z_k} = \sum_{i=1}^{n} \mathbb{1}_{Y_j \geq X_i} + \sum_{k=1}^{m} \mathbb{1}_{Y_j \geq Y_k} = \sum_{i=1}^{n} \mathbb{1}_{Y_j \geq X_i} + S_{Y,j}$$

où $(Z_1, ..., Z_{n+m})$ sont les $n+m$ données réunies, S_j est le rang de Y_j dans la suite $(Z_1, ..., Z_{n+m})$ et $S_{Y,j}$ est le rang de Y_j dans la suite $Y_1, ..., Y_m$.

- Somme des rangs

$$\sum_{j=1}^{m} S_j = \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}_{Y_j \geq X_i} + \sum_{j=1}^{m} S_{Y,j} = U_y + \frac{m(m+1)}{2}$$

Donc

$$W_y = U_y + \frac{m(m+1)}{2}.$$

Loi de U_y sous l'hypothèse H_0

- Loi exacte[a]

  U_y prend ses valeurs dans $\{0, 1, ..., nm\}$ et $P[U_y = k] \triangleq p_{n,m}(k)$ peut se calculer par récurrence sous H_0

$$p_{n,0}(k) = p_{0,m}(k) = \begin{cases} 1 \text{ si } k = 0 \\ 0 \text{ sinon} \end{cases}$$

  et pour $n \geq 1, m \geq 1$ et $k \in \{0, 1, ..., nm\}$

$$p_{n,m}(k) = \frac{n}{n+m} p_{n-1,m}(k-m) + \frac{m}{n+m} p_{n,m-1}(k).$$

[a] Sous Matlab, la loi exacte est utilisée pour $n + m < 20$ et $\min\{n, m\} < 10$.

Loi de U_y sous l'hypothèse H_0

- Loi asymptotique $(n, m \geq 8)$[a]

$$\frac{U_y - E[U_y]}{\sqrt{\text{var}[U_y]}} \underset{n,m\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0,1)$$

Attention, cette loi asymptotique est celle de U_y (ou de U_x) mais pas celle de $U = \min(U_x, U_y)$!! En présence d'ex aequo, on doit corriger la variance de U_y.

[a] G. Saporta, Probabilités, Analyse de données et Statistique, Editions Technip, Paris, France, p. 343, 2006.

Moyenne et variance de U_y sous H_0

- Moyenne

$$E[U_y] = \sum_{i=1}^{n}\sum_{j=1}^{m} E[\mathbb{1}_{Y_j > X_i}] = \frac{nm}{2}$$

- Variance

  En utilisant la notation $Z_{i,j} = \mathbb{1}_{Y_j > X_i}$, on a

$$\text{var}[U_y] = \text{var}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j}\right] = \sum_{i,j} \text{var}\left[Z_{i,j}\right] + \sum_{(i,j)\neq(i',j')} \text{cov}\left[Z_{i,j}, Z_{i',j'}\right] \tag{7}$$

$$= \frac{nm}{4} + \frac{nm(n+m-2)}{12} = \frac{nm(n+m+1)}{12}$$

- En présence d'ex aequos, la moyenne et la variance de U_y sont [Lehmann, p. 20]

$$E[U_y] = \frac{nm}{2},\ \text{var}[U_y] = \frac{nm(n+m+1)}{12} - \frac{nm\sum_{k=1}^{K}(d_k^3 - d_k)}{12(n+m)(n+m-1)}$$

où d_k est le nombre d'ex aequo pour le rang k et K est le nombre de rangs avec ex-aequos.

Preuves

$Z_{i,j}$ suit une loi de Bernoulli de moyenne $P[Y_j > X_i] = p = 1/2$ et de variance $p(1-p) = 1/4$

- Moyenne de U_y : $E[U_y] = nm \times \frac{1}{2} = \frac{nm}{2}$.
- Somme des variances de $Z_{i,j}$: $\sum_{i,j} \text{var}\,[Z_{i,j}] = \frac{nm}{4}$.
- Covariances
  - $i \neq i'$ et $j \neq j'$

    $$\text{cov}\left[Z_{i,j}, Z_{i',j'}\right] = 0.$$

  - $i = i'$ et $j \neq j'$: on considère les trois cas équiprobables $y_{j'} > y_j > x_i$, $y_{j'} > x_i > y_j$ et $x_i > y_{j'} > y_j$ (on peut supposer $y_{j'} > y_j$ sans perte de généralité) qui donnent respectivement $E[Z_{i,j}, Z_{i',j'}] = 1$, $E[Z_{i,j}, Z_{i',j'}] = 0$ et $E[Z_{i,j}, Z_{i',j'}] = 0$, d'où $E[Z_{i,j}, Z_{i',j'}] = \frac{1}{3}$ et $\text{cov}\left[Z_{i,j}, Z_{i',j'}\right] = \frac{1}{12}$.
  - $i \neq i'$ et $j = j'$: même résultat que ci-dessus
- Conclusion

  Comme il y a $nm(m-1)$ termes pour lesquels $i = i'$ et $j \neq j'$ et $mn(n-1)$ termes pour lesquels $i \neq i'$ et $j = j'$, on obtient

  $$\sum_{(i,j)\neq(i',j')} \text{cov}\left[Z_{i,j}, Z_{i',j'}\right] = nm(m-1) \times \frac{1}{12} + mn(n-1) \times \frac{1}{12} = \frac{nm(n+m-2)}{12}$$

Preuve de la récurrence

- $A_{n,m}(k)$

$$p_{n,m}(k) = \frac{A_{n,m}(k)}{\frac{(n+m)!}{n!m!}}$$

où $A_{n,m}(k)$ est le nombre de suites $z_1, ..., z_{n+m}$ possédant k "1" précédant des "0". Par exemple, si $\{x_1, x_2, x_3\} = \{2.1, 0.8, 5\}$ et $\{y_1, y_2\} = \{3, 5.8\}$, alors

$$y_2 = 5.8 > x_3 = 5 > y_1 = 3 > x_1 = 2.1 > x_2 = 0.8$$

d'où

$$z_1 = 1, z_2 = 0, z_3 = 1, z_4 = 0, z_5 = 0.$$

Comme z_1 précède 3 zéros et que z_3 précède 2 zéros, on a $U = 5$.

- Récurrence

$$A_{n,m}(k) = A_{n-1,m}(k-m) + A_{n,m-1}(k).$$

  - si $z_{n+m} = 0$, alors tous les y_j sont supérieurs à cet élément et donc $A(n,m) = k$ si et ssi $A(n-1,m) = k-m$
  - si $z_{n+m} = 1$, alors on peut enlever cet élément y_j de la suite et donc $A(n,m) = k$ si et ssi $A(n,m-1) = k$

Remarques

- Corrections
  - Continuité

$$U'_y = \frac{1}{E}\left[U_y - \frac{nm}{2} \pm 0.5\right] \sim \mathcal{N}(0,1) \text{ avec } E^2 = \frac{nm(n+m+1)}{12}$$

    avec $+0.5$ pour $F < G$ et -0.5 pour $F > G$.
  - Ex-aequos

    Remplacer le rang des ex-aequos par le rang moyen
- Pros
  - Le test marche bien pour des tailles d'échantillons faibles (taille minimale de chaque échantillon égale à 4)
  - Ne nécessite pas d'avoir d'info sur la loi des données à tester
  - C'est un des tests non-paramétriques les plus puissants
- Cons
  - Le test de Student (t-test) est en général plus puissant que le test de Mann-Whitney
  - Le test de Mann-Whitney peut détecter l'hypothèse alternative à tort

$p-$valeurs

- Test bilatéral ($G \neq F$)
  On rejette H_0 si $U_y < S_{1,\alpha}$ ou $U_y > S_{2,\alpha}$ avec une $p-$valeur définie par

$$\text{p-val} = 2\left[1 - F\left(\frac{|U_y - E[U_y]| - 0.5}{\sqrt{\text{var}[U_y]}}\right)\right]$$

- Test unilatéral à gauche ($F < G$)
  On rejette H_0 si $U_y < S_{1,\alpha}$ avec une $p-$valeur définie par

$$\text{p-val} = F\left(\frac{U_y - E[U_y] + 0.5}{\sqrt{\text{var}[U_y]}}\right)$$

- Test unilatéral à droite ($F > G$)
  On rejette H_0 si $U_y > S_{1,\alpha}$ avec une $p-$valeur définie par

$$\text{p-val} = 1 - F\left(\frac{U_y - E[U_y] - 0.5}{\sqrt{\text{var}[U_y]}}\right)$$

$p-$valeurs

Tables de $P[U \leq s]$ (https://www.cons-dev.org/elearning/stat/Tables/Tab6.html)

n2 = 7							
n1 / U	1	2	3	4	5	6	7
0	0,125	0,028	0,008	0,003	0,001	0,001	0,000
1	0,250	0,056	0,017	0,006	0,003	0,001	0,001
2	0,375	0,111	0,033	0,012	0,005	0,002	0,001
3	0,500	0,167	0,058	0,021	0,009	0,004	0,002
4	0,625	0,250	0,092	0,036	0,015	0,007	0,003
5		0,333	0,133	0,055	0,024	0,011	0,006
6		0,444	0,192	0,082	0,037	0,017	0,009
7		0,556	0,258	0,115	0,053	0,026	0,013
8			0,333	0,158	0,074	0,037	0,019
9			0,417	0,206	0,101	0,051	0,027
10			0,500	0,264	0,134	0,069	0,036
11			0,583	0,324	0,172	0,090	0,049
12				0,394	0,216	0,117	0,064
13				0,464	0,265	0,147	0,082
14				0,538	0,319	0,183	0,104
15					0,378	0,223	0,130
16					0,438	0,267	0,159
17					0,500	0,314	0,191
18					0,562	0,365	0,228
19						0,418	0,267
20						0,473	0,310
21						0,527	0,355
22							0,402
23							0,451
24							0,500
25							0,549

$p-$valeurs

Tables de $P[U \leq s]$ (https://www.cons-dev.org/elearning/stat/Tables/Tab6.html)

n2 = 8										
n1 U	1	2	3	4	5	6	7	8	t	Normal
0	0,111	0,022	0,006	0,002	0,001	0,000	0,000	0,000	3,308	0,001
1	0,222	0,044	0,012	0,004	0,002	0,001	0,000	0,000	3,203	0,001
2	0,333	0,089	0,024	0,008	0,003	0,001	0,001	0,000	3,098	0,001
3	0,444	0,133	0,042	0,014	0,005	0,002	0,001	0,001	2,993	0,001
4	0,556	0,200	0,067	0,024	0,009	0,004	0,002	0,001	2,888	0,002
5		0,267	0,097	0,036	0,015	0,006	0,003	0,001	2,783	0,003
6		0,356	0,139	0,055	0,023	0,010	0,005	0,002	2,678	0,004
7		0,444	0,188	0,077	0,033	0,015	0,007	0,003	2,573	0,005
8		0,556	0,248	0,107	0,047	0,021	0,010	0,005	2,468	0,007
9			0,315	0,141	0,064	0,030	0,014	0,007	2,363	0,009
10			0,387	0,184	0,085	0,041	0,020	0,010	2,258	0,012
11			0,461	0,230	0,111	0,054	0,027	0,014	2,153	0,016
12			0,539	0,285	0,142	0,071	0,036	0,019	2,048	0,020
13				0,341	0,177	0,091	0,047	0,025	1,943	0,026
14				0,404	0,217	0,114	0,060	0,032	1,838	0,033
15				0,467	0,262	0,141	0,076	0,041	1,733	0,041
16				0,533	0,311	0,172	0,095	0,052	1,628	0,052
17					0,362	0,207	0,116	0,065	1,523	0,064
18					0,416	0,245	0,140	0,080	1,418	0,078
19					0,472	0,286	0,168	0,097	1,313	0,094
20					0,528	0,331	0,198	0,117	1,208	0,113
21						0,377	0,232	0,139	1,102	0,135
22						0,426	0,268	0,164	0,998	0,159
23						0,475	0,306	0,191	0893	0,185
24						0,525	0,347	0,221	0,788	0,215
25							0,389	0,253	0,683	0,247
26							0,433	0,87	0,578	0,282
27							0,478	0,323	0,473	0,318
28							0,522	0,360	0,368	0,356
29								0,399	0,263	0,396
30								0,439	0,158	0,437
31								0,480	0,052	0,481
32								0,520		

Exemple

On souhaite comparer deux médicaments pour soulager la douleur post-opératoire. On a observé 16 patients, dont 8 ont pris le médicament A habituel, et les 8 autres un médicament B expérimental. Dans le tableau suivant sont reportés les temps (en heures) entre la prise du médicament et la sensation de soulagement. Effectuer un test de Mann-Whitney avec l'hypothèse $H_1 : G < F$ (temps avant soulagement plus grand avec médicament B qu'avec le médicament A) pour déterminer si le médicament A a un meilleur effet que le médicament B sur ces patients.

Médicament A	6.8	3.1	5.8	4.5	3.3	4.7	4.2	4.9
Médicament B	4.4	2.5	2.8	2.1	6.6	0.0	4.8	2.3

- Suite ordonnée

  $z(.) = (0.0, 2.1, 2.3, 2.5, 2.8, 3.1, 3.3, 4.2, 4.4, 4.5, 4.7, 4.8, 4.9, 5.8, 6.6, 6.8)$

- Médicaments associés

  (mB6, mB4, mB8, mB2, mB3, mA2, mA5, mA7, mB1, mA4, mA6, mB7, mA8, mA3, mB5, mA1)

- rangs du médicament B

  $r_1 = 9, r_2 = 4, r_3 = 5, r_4 = 2, r_5 = 15, r_6 = 1, r_7 = 12, r_8 = 3, W_y = \sum_{i=1}^{8} r_i = 51$

Exemple

- Statistique de Mann-Whitney
  - $U_y = W_y - \frac{8\times 9}{2} = 15.$
  - $U_x = W_x - \frac{8\times 9}{2} = nm - U_y = 64 - 15 = 49.$
  - $U = \min\{U_x, U_y\} = 15.$
- $p-$value : en utilisant l'approximation normale de la loi de Mann-Whitney avec correction de continuité, on obtient

$$p\text{-value} = 1 - F\left(\frac{U_y - E[U_y]}{\sigma_y}\right)$$

avec $E[U] = \frac{nm}{2} + 0.5$, $\sigma_y^2 = \frac{nm(n+m+1)}{12}$ et F est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$. Après application numérique, on obtient p-value ≈ 0.9584 (les tables donnent p-value ≈ 0.9590).

- Seuil pour $\alpha = 0.05$:

$$S_{\alpha,n,m} = E[U_y] + \sigma_y F^{-1}(0.95) \approx 47.16$$

Donc on accepte H_0 avec $\alpha = 0.05$ et on décide que B est aussi efficace que A.

Exemple

- Commande sous Matlab : [p,h,stats] = ranksum(x_B, x_A,'tail','right') car sous Matlab, la routine ranksum calcule les rangs de la première variable[a]. On peut aussi utiliser mwwtest(x_B,x_A) (avec $F_B > F_A$ sous H_1).

[a] pour des échantillons de même taille, sinon on calcule les rangs de l'échantillon de plus petite taille

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
- Chapitre 3 : Tests basés sur les rangs
  - Test de Wilcoxon-Mann Whitney
  - Test de la médiane
- Chapitre 4 : Tests de normalité

Test de la médiane

Le test de la médiane est un test non paramétrique qui permet de tester si deux échantillons indépendants $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ de fonctions de répartition F et G ont la même loi ou pas

$$H_0 : F = G, \quad H_1 : G < F$$

- Définition : on rejette H_0 si

$$M_{X,Y} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{R_j > \frac{n+m+1}{2}} > s_{n,m,\alpha}$$

où R_j est le rang de Y_j dans la suite constituée des $n + m$ données réunies $x_1, ..., x_n, y_1, ..., y_m$ [a] et $\frac{n+m+1}{2}$ est la médiane de cette suite.

[a] La plus petite de ces données a le rang 1, la suivante le rang 2, ... , la plus grande le rang $n + m$

Loi de $M_{X,Y}$ sous l'hypothèse H_0

- Loi exacte

  La variable aléatoire $mM_{X,Y}$ prend ses valeurs dans $\{0, 1, ..., m\}$ (nombre de y_j supérieurs à la médiane de la suite $x_1, ..., x_n, y_1, ..., y_m$) et possède une loi hypergéométrique.

- Rappel : Loi hypergéométrique

  La probabilité d'obtenir k boules blanches après n tirages sans remise dans une urne à deux catégories (constituée de b boules blanches et de r boules rouges avec $N = b + r$) notée P_k est définie par la loi hypergéométrique

$$P_k = \frac{\binom{b}{k}\binom{r}{n-k}}{\binom{N}{n}} = \frac{\binom{b}{k}\binom{N-b}{n-k}}{\binom{N}{n}}, \quad \forall k \in \{\max\{0, n-r\}, ..., \min\{b, n\}\}$$

- Remarque

  D'autres tests (comme le test de Kolmogorov Smirnov) sont généralement plus puissants que le test de la médiane qui pourrait être ignoré.

Que faut-il savoir ?

- Principe et mise en oeuvre d'un test de Mann-Whitney et avoir compris son lien avec le test de Wilcoxon
- Connaître l'existence et le principe du test de la médiane

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité
  - Droite de Henry
  - Test de Lilliefors
  - Test de Shapiro-Wilk

Droite de Henry

Si X est une variable aléatoire gaussienne de moyenne m et de variance σ^2, alors

$$P[X \leq x_i] = P\left[\frac{X-m}{\sigma} \leq \frac{x_i - m}{\sigma}\right] = P[Y \leq t_i] = \Phi(t_i)$$

où Y suit une loi normale centrée réduite (i.e., $Y \sim \mathcal{N}(0,1)$), Φ est la fonction de répartition de la loi normale centrée réduite et $t_i = \frac{x_i - m}{\sigma}$. Le principe de la droite de Henry est de tracer les paires $(x_{(i)}, t_{(i)})$ et de vérifier si les points obtenus sont alignés. Plus précisément

- 1) On détermine la statistique d'ordre de l'échantillon : $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
- 2) Pour tout $x_{(i)}$, on estime $P[X \leq x_{(i)}]$ avec la fonction de répartition empirique, i.e., $P[X \leq x_{(i)}] \approx \frac{i}{n}$.
- 3) On détermine $t_{(i)}$ par

$$t_{(i)} = \Phi^{-1}\left(\frac{i}{n}\right) \approx \Phi^{-1}\left[\int_{-\infty}^{x_{(i)}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(u-m)^2}{2\sigma^2}\right)\right] = \frac{x_{(i)} - m}{\sigma}$$

  où Φ est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.
- 4) On trace le Q-Q plot (Quantile-Quantile plot) constitué des points $(x_{(i)}, t_{(i)})$ et on vérifie s'ils sont alignés.

Exemples

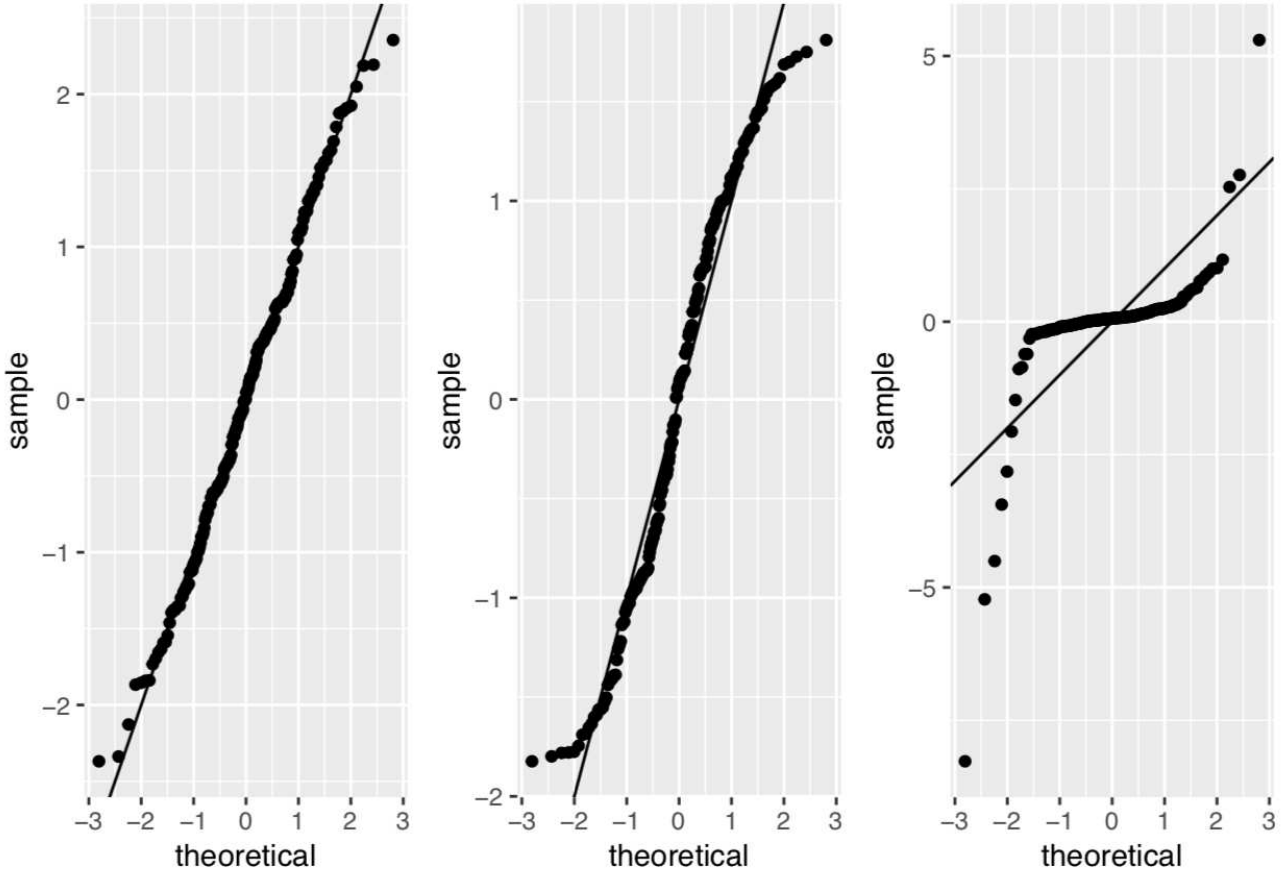


FIGURE 2.5 – Q-Q plot pour les 3 échantillons ($\mathcal{N}(2,1)$ à gauche, $\mathcal{U}([2,4])$ au centre, et $\mathcal{C}(1)$ à droite).

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité
  - Droite de Henry
  - Test de Lilliefors
  - Test de Shapiro-Wilk

Test de Lilliefors

Le test de Lilliefors est un test non paramétrique qui permet de tester si un échantillon $(X_1, ..., X_n)$ suit une loi normale ou pas

$$H_0 : X_i \sim \mathcal{N}(m, \sigma^2), \quad H_1 : \text{ non } H_0$$

- Définition

$$\text{Rejet de } H_0 \text{ si } D_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - \Phi(x; \bar{X}, S_n^2) \right| > S_{n,\alpha}$$

où $\Phi(x; m, \sigma^2)$ est la fonction de répartition d'une loi normale $\mathcal{N}(m, \sigma^2)$ et

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$$

Remarques

- Statistique de test

  C'est un test de Kolmogorov avec une fonction de répartition dans laquelle on a remplacé m et σ^2 par leurs estimateurs du maximum de vraisemblance, donc $D_n = \max_{i=1,\dots,n}\{E_i^+, E_i^-\}$

$$E_i^+ = \left|\Phi_{0,1}\left(\frac{x_{(i)} - \bar{x}}{s_n}\right) - \frac{i}{n}\right|, E_i^- = \left|\Phi_{0,1}\left(\frac{x_{(i)} - \bar{x}}{s_n}\right) - \frac{i-1}{n}\right|$$

  où $\Phi_{0,1}$ est la fonction de répartition d'une loi normale $\mathcal{N}(0,1)$.

- Loi de la statistique de test

  La loi de D_n ne dépend que de la loi de $\frac{X_{(i)} - \bar{X}}{S_n}$ qui ne dépend ni de m, ni de σ^2. La loi de D_n a été tabulée par Lilliefors[a]. Il existe plusieurs approximations conduisant par exemple pour $n \geq 50$ à

$$S_{n,0.05} = \frac{0.895}{f(n)} \text{ avec } f(n) = \frac{0.83 + n}{\sqrt{n}}.$$

[a] On peut par exemple approcher cette loi à l'aide de simulations de Monte Carlo avec des observations de la loi normale $\mathcal{N}(0,1)$.

Preuve

- Loi de D_n indépendante de m et de σ^2

  Si on pose $y_i = \frac{x_i - m}{\sigma}$, alors $x_{(i)} = \sigma y_{(i)} + m$, $\bar{x} = \sigma \bar{y} + m$ et donc

$$\frac{x_{(i)} - \bar{x}}{s_n} = \frac{\sigma(y_{(i)} - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}/\sqrt{n-1}} = \frac{y_{(i)} - \bar{y}}{\sqrt{\sum_i (y_i - \bar{y})^2}/\sqrt{n-1}}$$

  qui est une quantité indépendante de m et de σ^2 car $y_i \sim \mathcal{N}(0,1)$.

Plan du cours

- Chapitre 1 : Généralités sur les tests
- Chapitre 2 : Tests d'adéquation de Kolmogorov et du khi-deux
- Chapitre 3 : Tests basés sur les rangs
- Chapitre 4 : Tests de normalité
  - Droite de Henry
  - Test de Lilliefors
  - Test de Shapiro-Wilk

Test de Shapiro-Wilk

Le test de Shapiro-Wilk est un test non paramétrique qui permet de tester si un échantillon $(X_1, ..., X_n)$ suit une loi normale ou pas

$$H_0 : X_i \sim \mathcal{N}(m, \sigma^2), \quad H_1 : \text{ non } H_0$$

- Définition

$$\text{Rejet de } H_0 \text{ si } \mathsf{SW}_n = \frac{\left[\sum_{i=1}^n a_i X_{(i)}\right]^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} < S_{n,\alpha}$$

où $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $X_{(1)} < X_{(2)} < ... < X_{(n)}$ est la statistique d'ordre de l'échantillon $X_1, ..., X_n$. Les coefficients a_i sont tels que, à une constante multiplicative près, $\sum_{i=1}^n a_i X_{(i)}$ est le meilleur estimateur linéaire non biaisé de l'écart type des variables X_i dans le cas gaussien (hypothèse H_0).

Remarques

- Statistique de test
  - Interprétation

    W est le rapport de deux estimateurs de la variance σ^2 des variables X_i (à des constantes de normalisation près) obtenu sous hypothèse de normalité à partir de $(X_1, ..., X_n)$ et de $(X_{(1)}, ..., X_{(n)})$. On comprend donc que sous l'hypothèse H_0, W est proche de 1. Sous l'hypothèse alternative, Shapiro et Wilk ont montré que W prenait des valeurs inférieures à 1 (Cauchy-Schwartz).
  - Loi sous H_0

    La loi de la statistique de test SW_n sous l'hypothèse H_0 ne dépend que de n, ce qui permet de tabuler les seuils $S_{n,\alpha}$.
  - il y a un lien entre les coefficients a_i (à savoir $a_{n-i+1} = -a_i$) qui permet de réécrire la statistique de test comme suit

$$\mathrm{SW}_n = \frac{\left[\sum_{i=1}^{\mathrm{ent}(\frac{n}{2})} a_i (X_{(n-i+1)} - X_{(i)})\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}.$$

- Propriété

  Lorsque l'hypothèse H_1 est "$X_1, ..., X_n$ est un échantillon d'une loi continue qui n'est pas la loi normale", alors le test de Shapiro-Wilk est le test de normalité le plus puissant.

Expression des coefficients a_i

- Notations

  Soit $(X_1, ..., X_n)$ une suite de moyenne $E[X_i] = \mu$ et de variance var$(X_i) = \sigma^2$. La statistique d'ordre de la suite centrée réduite $(Y_1, ..., Y_n) = \left(\frac{X_1-\mu}{\sigma}, ..., \frac{X_n-\mu}{\sigma}\right)$ est notée $Y_{(1)}, ..., Y_{(n)}$ et possède un vecteur moyenne noté $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$ avec $\alpha_i = E[Y_{(i)}]$ et une matrice de covariance notée $\boldsymbol{B}$, qui ont une expression explicite sous l'hypothèse H_0.

- Estimation

  Pour estimer les coefficients a_i, on décompose les observations ordonnées $X_{(i)}$ comme suit

  $$X_{(i)} = \mu + \alpha_i \sigma + \epsilon_i$$

  et on cherche à estimer les paramètres μ et σ à l'aide de la méthode des moindres carrés pondérés (qui est aussi l'estimateur linéaire de variance minimale appelé estimateur BLUE ("Best linear unbiased estimator")).

Méthodes des moindres carrés pondérés

Comme $E[\epsilon_i] = E[X_{(i)}] - \mu - \alpha_i\sigma = 0$ (car $X_{(i)} = \sigma Y_{(i)} + \mu$ et $E[Y_{(i)}] = \alpha_i$) et que la matrice de covariance de $(\epsilon_1, ..., \epsilon_n)$ est celle de $(X_{(1)}, ..., X_{(n)})$ qui vaut $\sigma^2\boldsymbol{B}$ dans le cas gaussien, la méthode des moindres carrés pondérée consiste à rechercher les paramètres μ et σ qui minimisent

$$\left[\boldsymbol{X}_{(.)} - \boldsymbol{A}\begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right]^T \boldsymbol{B}^{-1}\left[\boldsymbol{X}_{(.)} - \boldsymbol{A}\begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right] \tag{8}$$

avec $\boldsymbol{X}_{(.)} = (X_{(1)}, ..., X_{(n)})^T$ et où $\boldsymbol{A}$ est une matrice de taille $n \times 2$ dont la première colonne contient des 1 et la seconde colonne contient le vecteur $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T$ (voir poly pour la résolution de ce problème des moindres carrés pondérés). Dans le cas où la loi des observations ordonnées est symétrique par rapport à sa moyenne, on obtient

$$\widehat{\mu} = \frac{\mathbf{1}^T\boldsymbol{B}^{-1}\boldsymbol{X}_{(.)}}{\mathbf{1}^T\boldsymbol{B}^{-1}\mathbf{1}} \quad \text{et} \quad \widehat{\sigma} = \frac{\boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{X}_{(.)}}{\boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha}} = \sum_{i=1}^{n}\gamma_i X_{(i)}$$

où $\boldsymbol{\alpha}$ et $\boldsymbol{B}$ sont le vecteur des moyennes et la matrice de covariance des statistiques d'ordre d'un échantillon de loi normale de taille n.

Méthodes des moindres carrés pondérés

La statistique du test de Shapiro-Wilk est alors définie par

$$\text{SW}_n = \frac{R^4 \widehat{\sigma}^2}{C^2[(n-1)S_n^2]} = \frac{\left[\sum_{i=1}^n a_i X_{(i)}\right]^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

avec

$$R^2 = \boldsymbol{\alpha}^T \boldsymbol{B}^{-1} \boldsymbol{\alpha}, C^2 = \boldsymbol{\alpha}^T \boldsymbol{B}^{-1} \boldsymbol{B}^{-1} \boldsymbol{\alpha}$$

et

$$\boldsymbol{a}^T = (a_1, ..., a_n) = \frac{\boldsymbol{\alpha}^T \boldsymbol{B}^{-1}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{B}^{-1} \boldsymbol{B}^{-1} \boldsymbol{\alpha}}}.$$

Remarques

- On a $\sum_{i=1}^n a_i^2 = \boldsymbol{a}^T \boldsymbol{a} = 1$, ce qui ne serait pas le cas si on avait considéré la statistique de test $\frac{\widehat{\sigma}^2}{S_n^2}$ (qui serait peut-être plus naturelle).
- Comme $E[\boldsymbol{X}_{(.)}] = \boldsymbol{\mu 1} + \sigma \boldsymbol{\alpha}$ et que $\boldsymbol{1}^T \boldsymbol{B}^{-1} \boldsymbol{\alpha} = 0$ pour une loi symétrique, les estimateurs $\widehat{\mu}$ et $\widehat{\sigma}$ sont clairement non biaisés.

Expression des coefficients a_i

Table 7 : Coefficients de Shapiro-Wilk :

n = taille de l'échantillon, i = numéro de la différence d_i

i / n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	0,7071														
3	0,7071	0													
4	0,6872	0,1677													
5	0,6646	0,2413	0												
6	0,6431	0,2806	0,0875												
7	0,6233	0,3031	0,1401	0											
8	0,6052	0,3164	0,1743	0,0561											
9	0,5888	0,3244	0,1976	0,0947	0										
10	0,5739	0,3291	0,2141	0,1224	0,0399										
11	0,5601	0,3315	0,226	0,1429	0,0695	0									
12	0,5475	0,3325	0,2347	0,1586	0,0922	0,0303									
13	0,5359	0,3325	0,2412	0,1707	0,1099	0,0539	0								
14	0,5251	0,3318	0,246	0,1802	0,124	0,0727	0,024								
15	0,515	0,3306	0,2495	0,1878	0,1353	0,088	0,0433	0							
16	0,5056	0,329	0,2521	0,1939	0,1447	0,1005	0,0593	0,0196							
17	0,4963	0,3273	0,254	0,1988	0,1524	0,1109	0,0725	0,0359	0						
18	0,4886	0,3253	0,2553	0,2027	0,1587	0,1197	0,0837	0,0496	0,0163						
19	0,4808	0,3232	0,2561	0,2059	0,1641	0,1271	0,0932	0,0612	0,0303	0					
20	0,4734	0,3211	0,2565	0,2085	0,1686	0,1334	0,1013	0,0711	0,0422	0,014					
21	0,4643	0,3185	0,2578	0,2119	0,1736	0,1399	0,1092	0,0804	0,053	0,0263	0				
22	0,459	0,3156	0,2571	0,2131	0,1764	0,1443	0,115	0,0878	0,0618	0,0368	0,0122				
23	0,4542	0,3126	0,2563	0,2139	0,1787	0,148	0,1201	0,0941	0,0696	0,0459	0,0228	0			
24	0,4493	0,3098	0,2554	0,2145	0,1807	0,1512	0,1245	0,0997	0,0764	0,0539	0,0321	0,0107			
25	0,445	0,3069	0,2543	0,2148	0,1822	0,1539	0,1283	0,1046	0,0823	0,061	0,0403	0,02	0		
26	0,4407	0,3043	0,2533	0,2151	0,1836	0,1563	0,1316	0,1089	0,0876	0,0672	0,0476	0,0284	0,0094		
27	0,4366	0,3018	0,2522	0,2152	0,1848	0,1584	0,1346	0,1128	0,0923	0,0728	0,054	0,0358	0,0178	0	
28	0,4328	0,2992	0,251	0,2151	0,1857	0,1601	0,1372	0,1162	0,0965	0,0778	0,0598	0,0424	0,0253	0,0084	
29	0,4291	0,2968	0,2499	0,215	0,1064	0,1616	0,1395	0,1192	0,1002	0,0822	0,065	0,0483	0,032	0,0159	0
30	0,4254	0,2944	0,2487	0,2148	0,187	0,163	0,1415	0,1219	0,1036	0,0862	0,0697	0,0537	0,0381	0,0227	0,0076

Table des valeurs limites de W

n	5%	1%
3	0.767	0.753
4	0.748	0.687
5	0.762	0.686
6	0.788	0.713
7	0.803	0.730
8	0.818	0.749
9	0.829	0.764
10	0.842	0.781
11	0.850	0.792
12	0.859	0.805
13	0.856	0.814
14	0.874	0.825
15	0.881	0.835
16	0.837	0.844
17	0.892	0.851
18	0.897	0.858
19	0.901	0.863
20	0.905	0.868
21	0.908	0.873
22	0.911	0.878
23	0.914	0.881
24	0.916	0.884
25	0.918	0.888
26	0.920	0.891
27	0.923	0.894
28	0.924	0.896
29	0.926	0.898
30	0.927	0.900

Que faut-il savoir ?

- Connaître et savoir mettre en oeuvre les principaux tests de normalité basés sur
  - une représentation graphique : droite de Henry
  - la fonction de répartition empirique : test de Lilliefors
  - les statistiques d'ordre : le test de Shapiro-Wilk

Synthèse

- Test de Kolmogorov

$$\text{Rejet de } H_0 \text{ si } D_n = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_0(x)| > S_{n,\alpha},$$

avec

$$D_n = \max_{i\in\{1,\ldots,n\}} \max\left\{ \left| \frac{i}{n} - F_0\left(x_{(i)}\right) \right|, \left| \frac{i-1}{n} - F_0\left(x_{(i)}\right) \right| \right\}$$

- Test de Kolmogorov-Smirnov

$$\text{Rejet de } H_0 \text{ si } D_{n,m} = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - \widehat{G}_m(x)| > S_{n,m,\alpha}$$

avec

$$D_{n,m} = \max_{j\in\{1,\ldots,n+m\}} \frac{n+m}{nm} \left| \frac{jm}{n+m} - \sum_{k=1}^{j} \alpha_k \right|$$

avec $\alpha_k = 1$ si la $k$ème plus petite observation appartient à la suite $\boldsymbol{y}$ et $\alpha_k = 0$ dans le cas contraire.

Synthèse

- Test du χ^2 d'ajustement

$$\text{Rejet de } H_0 \text{ si } \phi_n = \sum_{k=1}^{K} \frac{(Z_k - np_k)^2}{np_k} > S_{K,\alpha}$$

- Test du χ^2 d'adéquation à une loi

$$\text{Rejet de } H_0 \text{ si } \phi_n = \sum_{k=1}^{K} \frac{[Z_k - np_k(\hat{\theta})]^2}{np_k(\hat{\theta})} > S_{K,\alpha}$$

- Test du χ^2 d'indépendance et d'homogénéité

$$\text{Rejet de } H_0 \text{ si } I_n = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.} N_{.,l}}{n}\right)^2}{\frac{N_{k,.} N_{.,l}}{n}} > S_{K,L,\alpha}$$

Synthèse

- Test de Mann-Whitney

$$\text{Rejet de } H_0 \text{ si } U = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{Y_j > X_i} > S_{n,m,\alpha}$$

- Test de la médiane

$$\text{Rejet de } H_0 \text{ si } M_{X,Y} = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{R_j > \frac{n+m+1}{2}} > s_{n,m,\alpha}$$

où R_j est le rang de Y_j dans la suite constituée des $n+m$ données réunies $x_1, ..., x_n, y_1, ..., y_m$ [a] et $\frac{n+m+1}{2}$ est la médiane de cette suite.

[a] La plus petite de ces données a le rang 1, la suivante le rang 2, ... , la plus grande le rang $n+m$

Synthèse

- Test de normalité de Lilliefors

$$\text{Rejet de } H_0 \text{ si } D_n = \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - \Phi(x;\bar{X},S_n^2)\right| > S_{n,\alpha}$$

où $\Phi(x;m,\sigma^2)$ est la fonction de répartition d'une loi normale $\mathcal{N}(m,\sigma^2)$ et

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

- Test de Shapiro-Wilk

$$\text{Rejet de } H_0 \text{ si } \mathsf{SW}_n = \frac{\left[\sum_{i=1}^{n} a_i X_{(i)}\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} < S_{n,\alpha}$$

où $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ et $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ est la statistique d'ordre de l'échantillon $X_1, \ldots, X_n$. Les coefficients a_i sont tels que, à une constante multiplicative près, $\sum_{i=1}^{n} a_i X_{(i)}$ est le meilleur estimateur linéaire non biaisé de l'écart type des variables X_i.

Tests et machine learning

- Tests d'adéquation
  - Sélection de variables à l'aide d'un test de Kolmogorov-Smirnov [Biesiada, 2007].
  - Règle d'arrêt pour un arbre de décision à l'aide d'un test du χ^2 d'adéquation [Duda, 2000]
- Tests basés sur les rangs
  - Optimiser la performance d'un classifieur à deux classes
    Maximiser l'aire sous la courbe COR est équivalent à maximiser la statistique de Mann-Whitney SW entre les variables x_i associées à une classe et les variables y_j associées à l'autre classe, ce qui peut se faire à l'aide d'un algorithme de gradient appliqué à une approximation de SW [Yan, 2003].
- Tests de normalité et Mann-Whitney
  - Comparer la performance de plusieurs classifieurs
    Shapiro-Wilk est utilisé pour tester la normalité des différences des résultats de classification. Quand ces différences sont normales, les auteurs utilisent un test de Student, et dans le cas contraire un test de Mann-Withney pour déterminer si les résultats de classification sont différents ou pas [Jiang, 2016].

Références

- J. Biesiada and W. Duch, A Kolmogorov-Smirnov Correlation-Based Filter for Microarray Data, in Proc. Int. Conf. Neural Information Process. (ICONIP'07), Guangzhou, China, Nov. 2007.
- R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, 2nd edition, Wiley, 2000.
- L. Yan, R. Dodier, M. C. Mozer and R. Wolniewicz, Optimizing Classifier Performance via the Wilcoxon-Mann-Whitney Statistic, in Proc. Int. Conf. Machine Learning (ICML-03), Washington DC, USA, Aug. 2003.
- X. Jiang and D. L. Silver, Fishing Activity Detection from AIS Data Using Autoencoders, Canadian Conf. on Artificial Intelligence, Victoria, Canada, May 2016.