# Chapitre 3 : Tests Non Paramétriques

*Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses à la vue des résultats d'un échantillon. Il y a deux types de tests : paramétriques et non paramétriques. Le principal atout des tests non paramétriques est d'être à distribution free (libre), il n'est pas nécessaire de faire des hypothèses sur la forme des distributions.*

Soit $P$ l'ensemble deslois possibles pour une v.a. $X$. Supposons que $P$ ne dépend pas d'un paramètre (*distribution free). La distribution* $F_X$ sous $H_0$ et/ou $H_1$ n'est pas spécifiée. Un test NP dans ce cas est de type:

$$H_0 : \ F_X \sim Normale \ ?$$

Dans la suite, nous considérons deux v.a $X$ et $Y$ de lois $F_X$ et $F_Y$ resp. Soient $(X_1, ..., X_n)$, $(Y_1, ..., Y_m)$ deux échantillons issus de $X$ et de $Y$ resp. On cherche à tester

$$H_0 : F_X = F_Y \quad \text{contre} \quad H_1 : F_X \neq F_Y. \tag{1}$$

## I) Tests basés sur les rangs

### 1.1. Test de Wilcoxon

La majorité des tests non paramétriques reposent sur les rangs des observations. L'idée est substituer aux valeurs leur numéro dans l'ensemble des données. On étudie deux populations $P_1, P_2$ de deux variables qui représentent le même caractère quantitatif de loi continue. Elles sont notées : $X$ dans $P_1$ et $Y$ dans $P_2$. On veut copmarer (étudier l'homogénéité) les distributions de $X$ et de$Y$ :

$$\begin{cases} H_0 : \textit{les deux échantillons appartiennent à la même population} \\ H_1 : \textit{les deux échantillons sont de deux populations différents} \end{cases}$$

Notons $N = n + m$ la taille de l'échantillon global $(X_1, ..., X_n, Y_1, ..., Y_m)$.

Le test de **Wilcoxon (1945)**, consiste à ranger les observations $(X_1, ..., X_n, Y_1, ..., Y_m)$ par ordre croissant, il y'a donc $N! = (n + m)!$ ordres possibles. Notons par $R(X_i)$ le rang de $X_i$ $(i = 1, ..., n)$, le nombre possible de façons de placer les $X$ est:

$$\complement_n^N = \frac{(n+m)!}{n!m!}.$$

Soit la statistique de Wilcoxon des sommes des rangs :

$$W_X = \sum_{i=1}^{n} R(X_i) = \sum_{k=1}^{N} k\delta_k,$$

avec

$$\delta_k = \begin{cases} 1, & \text{si } X_i \text{ est en } k - ième \text{ position} \\ 0, & \text{sinon} \end{cases}$$

Alors,

$$E(W_X) = E(\delta_k) \sum_{k=1}^{N} k = \frac{n}{m} \frac{m(N+1)}{2} = \frac{n(N+1)}{2} \quad et \quad Var(W_X) = \frac{nm(N+1)}{12}.$$

### 1.2. Approximation par une loi normale

Lorsque les échantillons atteignent une taille suffisamment élevès ($n_1 > 8$ et $n_2 > 8$), la loi de la statistique $U$ converge vers la loi normale de moyenne $E(W)$ et de variance $V(W)$. Sous $H_0(F_X = F_Y)$, nous pouvons donc définir la statistique centrée réduite

$$Z = \frac{W - n(N+1)/2}{\sqrt{nm(N+1)/12}} \sim N(0,1)$$

La région critique du test au niveau de signification $\alpha$ est $|Z| > z_{1-\frac{\alpha}{2}}$, où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

**Remarque: Traitement des ex-aequos (principe des rangs moyens)** Quand on trouve des ex-aequons dans les valeurs, deux approches sont possibles. La méthode des rangs aléatoires attribue aléatoirement les rangs aux observations confondues. Dans ce cas, aucune modification des tables et lois asyptotiques existantes n'est pas nécessaire. Cependant, la puissance du test est faible que celle la méthode de traitement des ex-aequons. C'est pour cela la méthode des rangs moyens procède de la manière suivante : les observations possèdent des valeurs identiques se voient attribuer la moyenne de leurs rangs. Cette approche est plus puissantes que la précédente.

### 1.3. Test de Mann-Whitney

La procudure du **Mann-Whitney (1947)** est **:**

1. Classer toutes les observations par ordre croissant.
2. Affecter son rang à chaque observation.
3. Calculer $W$ la somme des rangs d'un échantillon (en général celui de plus petite taille).
4. Calculer les statistiques $U$ :

$$U_X = W_X - \frac{n(n+1)}{2}, \qquad U_Y = W_Y - \frac{m(m+1)}{2}, \quad U = \min(U_X, U_Y)$$

Lorsque l'hypothèse nulle est vraie, l'éspérance et la variance de $U$ s'écrivent

$$E(U) = \frac{nm}{2}, \quad Var(U) = \frac{nm(N+1)}{12}. \tag{2}$$

**Exemple :** Considérons deux groupes $X, Y$ de $n = 15$ enfants où les effets des maladies est "absent", et celui où les effets des maladies est "présent" avec $m = 12$. On chercher à tester l'hypothèse d'homogéniété des deux groupes ?

| groupes | *valeurs* | rang brut | rang moyen |
|---|---|---|---|
| *absent* | 6 | 1 | 1.5 |
| *present* | 6 | 2 | 1.5 |
| *absent* | 7 | 3 | 5 |
| *absent* | 7 | 4 | 5 |
| *absent* | 7 | 5 | 5 |
| *absent* | 7 | 6 | 5 |
| *absent* | 7 | 7 | 5 |
| *absent* | 8 | 8 | 9.5 |
| *absent* | 8 | 9 | 9.5 |
| *present* | 8 | 10 | 9.5 |
| *present* | 8 | 11 | 9.5 |
| *absent* | 9 | 12 | 12 |
| *absent* | 10 | 13 | 16 |
| *absent* | 10 | 14 | 16 |
| *absent* | 10 | 15 | 16 |
| *absent* | 10 | 16 | 16 |
| *present* | 10 | 17 | 16 |
| *present* | 10 | 18 | 16 |
| *present* | 10 | 19 | 16 |
| *present* | 11 | 20 | 20.5 |
| *present* | 11 | 21 | 20.5 |
| *absent* | 12 | 22 | 24.5 |
| *absent* | 12 | 23 | 24.5 |
| *present* | 12 | 24 | 24.5 |
| *present* | 12 | 25 | 24.5 |
| *present* | 12 | 26 | 24.5 |
| *present* | 12 | 27 | 24.5 |

| |
|---|
| *absent* |
| $n = 15$ |
| $W_X = \sum R(X_i) = 170.5$ |
| |
| |
| *present* |
| $m = 12$ |
| $W_Y = \sum R(X_i) = 207.5$ |
| |

On remarque que les observations ont été triées selon les valeurs croissantes. Un numéro global sert à repérer les individus. Il correspond aux rangs bruts. Il ne tient pas comte des ex-aequons puis, dans un deuxième temps, pour les observations ayant des valeurs identiques, nous attribuons la moyenne des rangs associés. Par exemple, les deux premières plus petites observations presentent la même valeur $x_1 = x_2 = 6$, nous leur attribuons le rang moyen $= 1.5$, pour les observations $x_3 = ... = x_7 = 7$, nous produisons le rang $= 5, etc...$

La somme et la moyenne des rangs conditionnellement aux groupe sont calculées:

$$U_X = 50.5, \qquad U_Y = 129.5, \quad U = \min(U_X, U_Y) = 50.5$$

Sous $H_0(F_X = Fy)$ :

$$\frac{U - nm/2}{\sqrt{nm(N+1)/12}} = \frac{50.5 - 90}{\sqrt{420}} = 1.93 < \ z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.05}{2}} = z_{0.975} = 1.96.$$

alors on acceptte $H_0$, (donc $F_X = F_Y$). On peut conclure que les deux groupes ont la même distribution.

### 1.4. Une autre version du test de Mann et Whitney

Le principe du test consiste à déterminer le nombre de couples $(X_i, Y_j)$ pour les quelles $X_i \neq Y_j$.

• Le test statistique de **Mann-Whitney** est défini par

$$U_{n,m} = \sum_{i=1}^{n}\sum_{j=1}^{m} 1_{(X_i > Y_j)}.$$

L'espérance et la variance de $U_{n,m}$ s'écrivent :

$$E(U_{n,m}) = \frac{nm}{2}, \qquad Var(U_{n,m}) = \frac{nm(N+1)}{12}.$$

Sous $H_0(F_X = Fy)$ :

$$\frac{U_{n,m} - nm/2}{\sqrt{nm(N+1)/12}} \sim N(0,1).$$

**Exemple .** Disposons de deux échantilllons (mâle et femelles) de souris des cactus (peromyscus eremicus), dont on a mesuré le poids (en g) chez l'individu adulte:

$$\begin{aligned} \textit{échantillon femelle } (n &= 6) : X = 24, 30, 30, 30, 38, 40 \\ \textit{échantillon mâle} (m &= 4) : Y = 20, 24, 26, 28 \end{aligned}$$

On veut tester si le poids des mâles et femelles sont les mêmes ou non ? Les hypothèses du test sont:

$$\begin{cases} H_0 : \textit{mâles et femelles ont le même poids.} \\ H_1 : \textit{mâles et femelles ont des poids différents.} \end{cases}$$

Pour plus de robustesse, on utilise le test de Mann-Whitney suivant la manière dont on calcul les rangs pour chaque échantillon. Ensuite, classant les couples $X, Y$ :

$$Z = 20, 24, 24, 26, 28, 30, 30, 30, 38, 40$$

puis on calcule $U_{n,m}$ :

$$U_{n,m} = 1 + 4 + 4 + 4 + 4 + 4 + 4 = 25.$$

Sous $H_0(F_X = Fy)$ :

$$U = \frac{U_{n,m} - nm/2}{\sqrt{nm(N+1)/12}} = \frac{25 - 6\,(4)\,/2}{\sqrt{6 \times 4(10+1)/12}} = 2.77 > z_{0.975} = 1.96.$$

Alors, on rejètte $H_0$, (donc $F_X \neq F_Y$) et en conclu que les poids des *mâles et des femelles ne sont plus les mêmes.*

### 1.5. Test de la médiane

Le problème (1) est équivalant au test de la médiane qui consiste à déterminer le nombre de variable $X$ qui sont strictement superieurs $(>)$ à la médiane de $(X, Y)$. Soit la statistique

$$M_{n,m} = \frac{1}{n}\sum_{j=1}^{n} 1_{\left(R(X_j) > \frac{N+1}{2}\right)}.$$

telle que $R(X_j)$ est le rang de la $j^{ème}$ observation de $X$.
* Sous $H_0$ et si $N = 2k$ $(pair)$:

$$E\left(M_{n,m}\right) = \frac{1}{2}, \qquad Var\left(M_{n,m}\right) = \frac{n}{4m\left(N-1\right)}.$$

* Sous $H_0$ et si $N = 2k+1$ $(impair)$:

$$E\left(M_{n,m}\right) = \frac{N-1}{2N}, \qquad Var\left(M_{n,m}\right) = \frac{n\left(N+1\right)}{2mN^2}.$$

**Remarque:** Sous $H_0$, la loi de $M_{n,m}$ est hypergeométrique, donc asympototiuqement normale (quand $\min\left(n, m\right) \to \infty$).

### 1.6. Test d'échelle

Supposons maintenant que $\forall x \in R : F_Y\left(x\right) = F_X\left(\frac{x}{\sigma}\right), \quad \sigma \in R^*$. Le problème (1) devient un test d'échelle :

$$H_0 : \sigma = 1 \ \text{ contre } H_1 : \sigma \neq 1.$$

Soient

$$A_{(1)} = \frac{1}{N}, \quad A_{(2)} = \frac{1}{N} + \frac{1}{N-1}, ..., A_{(k)} = \frac{1}{N} + \frac{1}{N-1} + ... + \frac{1}{N-k+1}, \qquad k = 1, ..., n$$

Le test d'échelle est défini par la statistique:

$$S = \sum_{j=1}^{n} A_{(R(X_j))}.$$

Sous $H_0 : S$ suit une loi normale $N\left(\mu, \sigma^2\right)$,

$$\mu = E\left(S\right) = n \quad \text{ et } \quad \sigma^2 = Var\left(S\right) = \frac{nm}{N-1}\left(1 - \frac{1}{N}\sum_{h=1}^{N}\frac{1}{h}\right).$$

**Exemple :** Appliquer le test d'échelle aux données de l'exemple précident de souris des cactus.

# II) Tests basés sur la distribution

Dans cette partie, trois types de tests basés sur les distributions (théorique et empirique) sont rencontrés :

- Test d'ajustement (Adéquation),
- Test d'homogénéité,
- Test d'indépendance.

## 2.1. Test et distance de Pearson - Khi deux

Soit $X_1, ..., X_n$, un éhantillon de $X$ de loi $P$ à valeurs dans un ensemble $O \in \mathbb{R}$, et soit $\{O_1, ..., O_m\}$ une partition de $O$, telle que:

$$O = \bigcup_{k=1}^{m} O_k, \quad O_j \cap O_k = \phi, \quad j \neq k$$

On définit le nombre $N_k$ de $X_i$ appartenant à $O_k$ par:

$$N_k = \sum_{i=1}^{N} 1_{(X_i \in O_k)} \qquad \forall k = 1, ..., m$$

Soient $p_1, .., p_m$, les probabilités pour les quelles : $p_k = P(X_i \in O_k)$. Alors, le veteur $(N_1, ..., N_m)$ suit une loi multinomiale $M(n, p_1, ..., p_m)$ :

$$P(N_1 = n_1, ..., N_m = n_m) = \frac{n!}{n_1!...n_m!} p_1^{n_1} ... p_m^{n_m}.$$

Posons aussi $P_n$ la loi empirique estimant $P$ sur la base de l'échantillon $(X_1, ..., X_n)$ par :

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

où $\delta_*$ est la masse de Dirac.

**Définition:** La distane de Khi-2 (*Pearson*, 1900) entre $P$ et $P_n$ est

$$Q = D(P, P_n) = \sum_{k=1}^{m} \frac{(N_n - nP_k)^2}{nP_k}, \quad avec \quad Q \sim \chi^2_{(m-1)} \quad quand \quad n \to \infty.$$

### 2.1.1. Test d'ajustement (adéquation) du khi-deux

À partir de l'échantion $X = (X_1, ..., X_n)$ on peut vérifié la qualité d'ajustement à une distribution théorique spécifiée par l'hypothèse nulle $H_0$. Posons $P_0$ une loi donnée et considérons le problème du test :

$$\begin{cases} H_0 : (P = P_0) \\ H_1 : (P \neq P_0) \end{cases}, \qquad \text{où } P_0(\theta_k) = P_{0k}.$$

Intuitivement, si les $X_i$ suivent la loi $P_0$, la distance de khi-2 $D(P_n, P_0)$ entre $P_n$ et $P_0$ sera petite ($Q$ décroit vers 0), par ailleurs on sait que si les $X_i$ suivent la loi $P_0$, alors $D(P_n, P_0)$ suit asymptotiquement une loi du $\chi^2$ à $(m-1)$ degrés de liberté.

- La statistique de khi-2 définie par :

$$Q = D(P_n, P_0) = \frac{\sum (N_k - np_{0k})^2}{np_{0k}}. \tag{3}$$

- La région critique : on rejette $H_0$ si

$$Q > \chi^2_{(m-1)}(1 - \alpha) = q_\alpha.$$

**Remarque:** Pour tester $H_0(P = P_{0,\theta})$, où $P_{0,\theta}$ est une famille de loi ($\theta \in \mathbb{R}^+$), alors $H_0$ est rejetté si $Q > \chi^2_{(m-r-1)}(1 - \alpha) = q_{\alpha r}$.

**Exercice:** Soient $n = 20$ observations d'une varaiable aléatoire $X$ de loi inconnue $P$ :

$$\begin{array}{llllllllll} -4.2, & -1.6, & -1.6, & -1.3, & -0.5, & -0.2, & -0.1, & 0.2, & 0.3 \\ 0.7, & 0.9, & 1.6, & 1.9, & 2.4, & 2.5, & 2.5, & 2.6, & 2.9, & 3.0, & 3.4 \end{array}$$

Notons $\{O_1, ..., O_4\}$ la partition de $R$, telle que:

$$O_1 = ]-\infty, -0.6], \quad O_2 = ]-0.6, 0.6], \quad O_1 = ]0.6, 2.5] \quad \text{et} \quad O_1 = ]2.5, \infty[.$$

Tester en utilisant le test de khi-deux l'hypothèse de normalité de $X$ : $H_0 : P \sim N\left(\mu, \sigma^2\right)$.

### 2.1.2. Test de khi-deux d'indépendance

Ce test est utilisé pour étudier sur un même échantillon de taille $n$ la liaison entre deux variables quantitatives. Soient $X, Y$ deux variables qualitatives telle que $X$ est à valeur dans $\{a_1, ..., a_m\}$ et $Y$ à valeur dans $\{b_1, ..., b_n\}$. Sous $H_0$, la distribution de $X$ devrait *être indépendante* de celle de $Y$. Par contre, si la disrtibution de $X$ est *liée* à celle de $Y$, on rejette $H_0$ au profit de $H_1$, les deux variables $X$ et $Y$ sont liées. Ainsi, on cherche à tester :

$$\begin{cases} H_0 : \text{Les variables } X, Y \text{ sont indépendantes} \\ H_1 : \text{Les variables sont liées} \end{cases}$$

• La statistique de khi-deux d'indépendance est :

$$Q_{ind} = \sum_{i=1}^{m}\sum_{j=1}^{l} \frac{(\frac{N_{i*}*N_{*j}}{n} - N_{ij})^2}{\frac{N_{i*}*N_{*j}}{n}}, \tag{4}$$

$N_{i*}$ : Nombre de $X$ de valeurs $a_i$ $(i = 1, ..., m)$.
$N_{*j}$ : Nombre de $Y$ de valeurs $b_j$ $(j = 1, ..., k)$.
$N_{ij}$ : Nombre de $(X, Y)$ de valeurs $(a_i, b_j)$.

• La région critique :pour un risque de première espèce $\alpha$, on rejette $H_0$ si

$$Q_{ind} > \chi^2_{(m-1)(p-1)}(1 - \alpha).$$

Autrement dit, si la valeur de la statistique de test $\chi^2$ est supérieur à la valeur du seuil $\chi^2_{(m-1)(p-1)}(1 - \alpha)$ alors on rejette l'hypothèse nulle, il existe donc une liaison significatif entre $X, Y$.

**Exemple:** On veux savoir est-ce que il y a une liaison entre les notes des étudiants et leurs sexe (fille, garçon). On prend les notes de 69 étudiants selon les classes (moyennes) de sexe (fille, garçon) avec un risque $\alpha = 0.05$.

| *notes* | $[0, 9[$ | $[9, 15[$ | $[15, 20[$ |
|---|---|---|---|
| $F$ | 25 | 14 | 15 |
| $G$ | 8 | 5 | 7 |

Sous $R$, on utilisons les commandes associées, pour $(k = 2, m = 3)$

$> A = matrix(c(25, 14, 15, 8, 5, 7), nrow = 2, byrow = T)$
$> chisq.test(A)\$p.value$
[1] 0.8225567
$p - valeur = 0.8225567$

On observons q'aucun "warning message" n'apparait, alors les conditions d'applications du test sont vérifiées, comme $p - valeur > 0.05$, on ne rejette pas $H_0$. Les données ne nous permettent pas de rejetter l'indépendane entre les notes et le sexe.

### 2.1.3. Test du khi-deux d'homogénéité

Le test du khi-deux peut utilisé aussi dans le cadre de la comparesion entre les lois (disributions) de deux échantillons indépendantes. Considérons un couple de $v.a$ $(Y, Z)$ à valeurs dans $\{a_1, ..., a_m\}$. Soient $Y_1, ..., Y_{n_1}, Z_1, ..., Z_{n_2}$ deux éhantillons de $Y$ et $Z$ respectivement. On cherche à tester l'homogénéité des lois de $Y$ et $Z$:

$$\begin{cases} H_0 : Y \textit{ et } Z \text{ ont la même loi.} \\ H_1 : Y \textit{ et } Z \text{ de loi différente.} \end{cases}$$

et on dispose de l'observation : $u = (y_1, ..., y_{n_1}, z_1, ..., z_{n_2})$ de $U = (Y_1, ..., Y_{n_1}, Z_1, ..., Z_{n_2})$.

• Statistique de test est :

$$\mathbf{Q}_{\text{hom}} = \sum_{k=1}^{m} \mathbf{n}_1 \frac{(\frac{N_k+M_k}{n_1+n_2} - \frac{N_k}{n_1})^2}{\frac{N_k+M_k}{n_1+n_2}} + \mathbf{n}_2 \frac{(\frac{N_k+M_k}{n_1+n_2} - \frac{M_k}{n_2})^2}{\frac{N_k+M_k}{n_1+n_2}}$$

telle que :

$N_k$ : le nombre d'éléments de $\{Y_1, ..., Y_{n_1}\}$ qui prennent la valeur $a_k$.

$M_k$ : le nombre d'éléments de $\{Z_1, ..., Z_{n_2}\}$ qui prennent la valeur $a_{k.}$

• On rejette $H_0$ si la valeur de $Q_{\text{hom}} > s$ où le choix de la valeur critique $s$, sous $H_0$, $Q_{\text{hom}}$ asymptotiquement une loi $\chi^2$ de $(m-1)$ dégré de liberté. Donc $s$ srea choisie telle que $P(k \geq s)$,avec $k \sim \chi^2(m-1)$.

## 2.2. Test et distance de Kolmogorov-Smirnov

L'idée du test est de *comparer la fonction de disribution empirique à la foncction de répartition*. Soit $X_1, ..., X_n$ un $n$-échantillon d'une va $X$ de loi $P$ **absolument continue** par rapport à la mesure de Lebesgue sur $(\mathbb{R}, \mathbf{B}(\mathbb{R}))$ inconnue. On note $F_n$ la distribution empirique associée à $X$.

·*Le théorème de Glivento − Gantelli* donne :

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \underset{n\to+\infty}{\longrightarrow} 0 \quad Ps. \tag{5}$$

· la statistique de K-S est défini par la distance en norme infinie de la fonction de répartition empirique $F_n$, et la fonction de répartition $F$ :

$$KS = D_{KS}(P, P_0) = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| . \tag{6}$$

**Proposition:** Si $(x_{(1)}, ..., x_{(n)})$ est la statistique d'ordre associée à l'échantillon $X$ alors :

$$D_{KS}(P, P_n) = \max_{1\leq i\leq n} \max \left\{ \left| F(x_{(i)}) - \frac{i}{n} \right|, \left| F(x_{(i)}) - \frac{i-1}{n} \right| \right\}.$$

On rejette $H_0$ si $\sqrt{n}D_{KS} > d_{n,\alpha}$, où $d_{n,\alpha}$ est le quantile théorique lu à partir la table de Kolmogorov-Smirnov.

**Exemple:** On souhaite étudier le temps $X$ (en mois) mais par 10 étudiants (diplomés) pour obtenir un emplois. On prend $3.5, 16, 18, 14, 26, 17.5, 12, 22.5, 36, 10$. On cherche à tester $H_0(X \sim Exp(\lambda = 1/5))$ avec un risque $\alpha = 0.05$.

Sous $R$, on utilisons la commande *ks.test* du package'*starts*' comme suit :
$X < -c(3.5, 16, 18, 14, 26, 17.5, 12, 22.5, 36, 10)$
$ks.test(X, "ppois", lambda = 1/5)$
$One - sample\ Kolmogorov - Smirnov\ test$
$data : X$
$D = 0.88248, p - value = 3.442e - 07$
$alternative\ hypothes\ is : two - sided$
$p - value = 0.003$
Comme la p-valeur est inférieure à la valeur de $\alpha$, alors on peut rejetter l'hypothèse nulle, c'est à dire accepter $H_1$. Danc la distribution observée ne suive pas la loi expononsielle de paramètre $1/5$ au risque 5%.

### 2.2.1. Test de Lilliefors

Ce test est une variante du test de Kolmogorov-Smirnov, sous l'hypothèse de normalité (à chercher à tester $H_0 : P \sim Gaussienne$), où les paramètres $\mu, \sigma$ de la loi sont estimés à partir des données.

• La statistique du test est :

$$L_n = \sqrt{n} \max_{1\leq i\leq n} \max \left\{ \left| F_0(\frac{x_{(i)} - \overline{x}}{S_x}) - \frac{i}{n} \right|, \left| F_0(\frac{x_{(i)} - \overline{x}}{S_x}) - \frac{i-1}{n} \right| \right\}$$

où $\overline{x}$ est la moyenne empirique et $S_x$ est l'écart type empirique.

• On rejette $H_0$ si $L_n > D_{crit}$ ($D_{crit}$ la valeur critique de test Lilliefors).

Sous R, après avoir charger le package de la fonction *lillie.test* de library (*nortest*), on peut utiliser la $p - value$ pour conclur l'acceptation de $H_0$ comme suit :

$> lillie.test(rnorm(100, mean = 5, sd = 3))$

$lilliefors(kolmogorov - smirnov)normality.test$
$data : rnorm(100, mean = 5, sd = 3)$
$D = 0.0646$, $p - value = 0.3841$.
$p - value = 0.3841 > 0.05$, donc $H_0$ est acceptée ($P \sim N(5, 9)$).

### 2.2.2. Test de Komogorov-Smirnov d'homogénéité

L'objetif de ce test est si l'on veut tester si les deux échantillons peuvent provenir de la même population. En d'autre terme **tester l'identiquité de deux distributions empiriques** à partir de deux $v\ a$ indépendantes $X$ et$Y$ de tailles $n_1$,$n_2$ respectivements et de *lois inconnues* et consédirons $F_1$, $F_2$ étant leurs fontions de répartitions et $F_{1_{n_1}}$, $F_{2_{n_2}}$ leurs fonctions de répartitions empiriques :

$$F_{1_n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(x_{(i)} \leq t)}, \qquad F_{2_m} = \frac{1}{m}\sum_{i=1}^{m} 1_{(y_{(i)} \leq t)}$$

On cherche à tester $H_0 : F_1(x) = F_2(x)$ contre $H_1 : F_1(x) \neq F_2(x)$. La distance de **K-S** d'homogénéité est :

$$D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup |F_{1_n}(t) - F_{2_m}(t)|$$

Le test de **K-S** d'homogénéité repose sur **l'écart maximum** entre les fonctions de répartitions empiriques.

- On rejette $H_0$ si $D_{n,m} > s_{KS}$.