Notions préliminaires

1 Introduction :

On qualifie de non paramétrique les méthodes statistiques applicables quelle que soit la loi mère (méthodes n'impliquant pas de spécification à priori de lois théoriques dépendant d'un nombre fini de paramètres). Les méthodes non paramétriques (distribution free) donnent des résultats indépendants de la forme des distributions théoriques des variables analysées, pourvu que celles-ci appartiennent à des familles assez générales (par exemple fonction de répartition continue). Les méthodes non paramétriques s'appliquent aux variables quantitatives mais aussi aux variables catégorielles (ordinales ou nominales).

Les tests non paramétriques ont de nombreux avantages pratiques :

- Ils font appel à des hypothèses simples et générales,
- Pour les petits échantillons (TCL non applicable) ils occasionnent des calculs simples,
- Ces tests sont parfois les seuls disponibles (les données de départ sont des classements par exemple).

Les statistiques d'ordre associées aux statistiques de rang font partie des outils fondamentaux de la statistique non paramétrique.

2 Fonction de répartition empirique :

2.1 Définition :

Soit $(X_1, \dots, X_n)$ un n-échantillon d'une v.a. X de f.r. F. La fonction de répartition empirique F_n est définie pour tout $x \in \mathbb{R}$ par

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

Notons que F_n est une fonction aléatoire et que pour tout $x \in \mathbb{R}$, $F_n(x)$ est une v.a.

2.2 Propriétés de la fonction de répartition empirique (propriétés ponctuelles) :

Pour tout $x \in \mathbb{R}$, on a

1. $F_n(x)$ est un estimateur sans biais de $F(x)$: $E(F_n(x)) = F(x)$,
2. $Var\big(F_n(x)\big) = \frac{F(x)(1-F(x))}{n}$,
3. Erreur en moyenne quadratique : $E((F_n(x) - F(x))^2) = Var\big(F_n(x)\big)$,
4. $F_n(x)$ converge en moyenne quadratique vers $F(x)$: $\lim_n E((F_n(x) - F(x))^2) = 0$,
5. $F_n(x)$ converge en probabilité vers $F(x)$: $\forall \varepsilon > 0, \lim_n P\{|F_n(x) - F(x)| > \varepsilon\} = 0$,
6. $F_n(x)$ converge presque surement vers $F(x)$: $P\left\{\lim_n F_n(x) = F(x)\right\} = 1$,

7. $\sqrt{n}(F_n(x) - F(x))$ converge en loi vers une v.a. de loi $N(0, F(x)(1-F(x))$,
8. Loi du logarithme itéré: $\lim_n sup \frac{\sqrt{n}|F_n(x)-F(x)|}{\sqrt{F(x)(1-F(x))2loglogn}} = 1 \quad P.S.$

Remarque :

Il existe plusieurs versions de la loi du log-itéré.

2.3 Propriétés uniformes :

1. Théorème de Glivenko Cantelli : $D_n = \sup_{x\in\mathbb{R}}|F_n(x) - F(x)|$ converge presque surement vers 0 (F_n converge uniformément vers F presque surement).
2. Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW) :
$$\forall n \in \mathbb{R}, \forall \varepsilon > 0, P\{D_n > \varepsilon\} \leq 2e^{-2n\varepsilon^2}.$$
3. Théorème de Kolmogorov:
$$\lim_n P\left(\sqrt{n}\, D_n \leq y\right) = K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k exp(-2k^2y^2).$$

Remarque :

La loi exacte de D_n est tabulée.

3 Statistiques d'ordre

3.1 Définition:

Soit $(X_1, \dots, X_n)$ un n-échantillon d'une variable aléatoire X de fonction de répartition F (et de densité f). L'échantillon ordonné dans l'ordre croissant noté $\left(X_{(1)}, \dots, X_{(n)}\right)$ est défini tel que $X_{(k)}$ soit la k *ième* plus petite valeur de l'échantillon $(X_1, \dots, X_n)$: $X_{(1)} \leq X_{(2)\leq} \dots \leq X_{(n)}$.

On appelle $X_{(k)}$ la statistique d'ordre k (ou la k *ième* statistique d'ordre).

Deux cas importants de statistiques d'ordre sont le minimum et le maximum, appelées valeurs extrêmes:

$$X_{(1)} = Min_{1\leq i\leq n}X_i,\ X_{(n)} = Max_{1\leq i\leq n}X_i.$$

Des fonctions importantes des statistiques d'ordre sont :

- La médiane (plus généralement les quantiles),
- L'étendue définie comme l'écart entre les deux valeurs extrêmes : $W = X_{(n)} - X_{(1)}$,
- La moyenne extrême : $\frac{X_{(1)}+X_{(n)}}{2}$,
- Les L-statistiques qui sont des combinaisons linéaires des statistiques d'ordre : $\sum_{i=1}^{n} \alpha_i X_{(i)}$,
- La fonction de répartition empirique

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_{(i)} \le x\}} = \begin{cases} 0 & si\ x < X_{(1)}, \\ \frac{1}{n} & si\ X_{(1)} \le x < X_{(2)}, \\ \vdots & \vdots \\ \frac{k}{n} & si\ X_{(k)} \le x < X_{(k+1)}, \\ \vdots & \vdots \\ 1 & si\ x \ge X_{(n)}. \end{cases}$$

Remarque :

Les statistiques d'ordre $X_{(k)}, k = \overline{1,n}$, ne sont pas indépendantes.

3.2 Fonction de répartition empirique inverse (Fonction quantile) :

Soit $(X_1, \dots, X_n)$ un n-échantillon de X et $\left(X_{(1)}, \dots, X_{(n)}\right)$ l'échantillon ordonné. La fonction quantile, notée F_n^{-1}, est définie sur $[0,1]$ par :

$$F_n^{-1}(p) = \inf\{t \in \mathbb{R},\ F_n(t) \ge p\}.$$

On obtient

$$F_n^{-1}(p) = X_{(i)} \quad si \quad \frac{i-1}{n} < p < \frac{i}{n}.$$

La valeur $F_n^{-1}(p)$ est aussi appelée le quantile empirique d'ordre p (noté Q_p).

Pour $p = \frac{1}{2}$, on obtient la médiane empirique $F_n^{-1}\left(\frac{1}{2}\right) = Q_{\frac{1}{2}}$

3.3 Loi d'une statistique d'ordre

Soit H_k la fonction de répartition de $X_{(k)}, k = \overline{1,n}$.

Théorème :

La fonction de répartition de $X_{(k)}$ est donnée par

$$H_k(x) = \sum_{j=k}^{n} C_n^j (F(x))^j (1-F(x))^{n-j}.$$

Si F possède une densité f alors la densité de $X_{(k)}$, notée h_k, est donnée par

$$h_k(x) = nC_{n-1}^{k-1}\big(F(x)\big)^{k-1}\big(1-F(x)\big)^{n-k} f(x) = kC_n^k\big(F(x)\big)^{k-1}\big(1-F(x)\big)^{n-k} f(x).$$

En particulier :

$$H_n(x) = (F(x))^n, \qquad h_n(x) = n\big(F(x)\big)^{n-1} f(x),$$

$$H_1(x) = 1-(1-F(x))^n, \qquad h_1(x) = n\big(1-F(x)\big)^{n-1} f(x).$$

Preuve :

Soit $R_n(x)$ le nombre de variable $X_i, 1 \le i \le n$, satisfaisant $\{X_i \le x\}$.

On a: $\{X_{(k)} \le x\} \Leftrightarrow \{R_n(x) \ge k\}$

Comme la variable aléatoire $R_n(x)$ suit une loi binomiale de paramètres n, $F(x)$, alors

$$H_k(x) = P\big(X_{(k)} \le x\big) = P(R_n(x) \ge k) = \sum_{j=k}^{n} C_n^j (F(x))^j (1-F(x))^{n-j}.$$

$$\left(= \frac{n!}{(k-1)!\,(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt \right)$$

Si X est absolument continue de densité f alors $X_{(k)}$ est absolument continue de densité h_k telle que :

$h_k(x) = H_k^{'}(x)$

$= \sum_{j=k}^{n} C_n^j \left\{ j\big(F(x)\big)^{j-1}(1-F(x))^{n-j} - (n-j)\big(F(x)\big)^{j}\big(1-F(x)\big)^{n-j-1} \right\} f(x)$

Les termes de la somme s'éliminent deux à deux. Après simplification, on obtient

$$h_k(x) = k C_n^k \big(F(x)\big)^{k-1}\big(1-F(x)\big)^{n-k} f(x).$$

Exemples :

1. $(X_1, \dots, X_n)$ un n-échantillon d'une loi exponentielle de paramètre λ :

$$h_n(x) = n\,\lambda\big(1-e^{-\lambda x}\big)^{n-1} e^{-\lambda x} \mathbb{I}_{\{x>0\}},$$
$$h_1(x) = n\,\lambda\, e^{-\lambda n x} \mathbb{I}_{\{x>0\}},\ X_{(1)} \rightsquigarrow \xi(n\lambda)\,.$$

2. $(X_1, \dots, X_n)$ un n-échantillon d'une loi uniforme sur $[0,1]$:

$$h_1(x) = n\,(1-x)^{n-1}\mathbb{1}_{[0,1]}(x), \quad X_{(1)} \rightsquigarrow \mathcal{B}_{[0,1]}(1,n),$$
$$h_n(x) = n\,x^{n-1}\mathbb{1}_{[0,1]}(x), \quad X_{(n)} \rightsquigarrow \mathcal{B}_{[0,1]}(n,1).$$

Remarque :

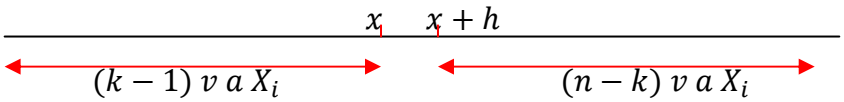
On a

$$h_k(x) = H_k^{'}(x) = \lim_{h\to 0} \frac{H_k(x+h) - H_k(x)}{h} = \lim_{h\to 0} \frac{P(x < X_{(k)} \le x+h)}{h}.$$

h étant très petit alors $(x + h \approx x)$.

L'événement $\{x < X_{(k)} \le x+h\}$ est équivalent à

$$\{(k-1)\ v.a\ X_i \le x, \quad x < 1\ v.a\ X_i \le x+h, \quad (n-k) v.a\ X_i > x+h\}$$



D'où

$$P\big(x < X_{(k)} \le x+h\big) = C_n^{k-1} C_{n-k+1}^1 C_{n-k}^{n-k} \big(F(x)\big)^{k-1}\big(F(x+h)-F(x)\big)\big(1-F(x+h)\big)^{n-k}$$

$$= \frac{n!}{(k-1)!\,(n-k)!} F(x)^{k-1}\big(F(x+h)-F(x)\big)\big(1-F(x+h)\big)^{n-k}.$$

Par suite:

$$h_k(x) = \frac{n!}{(k-1)!\,(n-k)!}(F(x))^{k-1}\lim_{h\to 0}\frac{\big(F(x+h)-F(x)\big)}{h}\lim_{h\to 0}\big(1-F(x+h)\big)^{n-k}$$

$$= \frac{n!}{(k-1)!\,(n-k)!}\big(F(x)\big)^{k-1} f(x)\big(1-F(x)\big)^{n-k}.$$

3.4 Loi conjointe de deux statistiques d'ordre :

Soit le couple $\big(X_{(i)}, X_{(j)}\big)$ de fonction de répartition conjointe $H_{i,j}$ et de densité conjointe $h_{i,j}$, $1 \le i < j \le n$.

Théorème :

$$H_{i,j}(x,y) = \begin{cases} H_j(y) & si\ x \ge y, \\ \sum_{k=j}^{n}\sum_{l=i}^{k} \frac{n!}{l!\,(k-l)!\,(n-k)!} F(x)^l\big(F(y)-F(x)\big)^{k-l}\big(1-F(y)\big)^{n-k} & si\ x < y, \end{cases}$$

$h_{i,j}(x,y) = \frac{n!}{(i-1)!(j-i-l)!(n-j)!} F(x)^{i-1}\big(F(y)-F(x)\big)^{j-i-l}\big(1-F(y)\big)^{n-j} f(x)f(y)\mathbb{1}_{\{x<y\}}$.

En particulier :

$$H_{1,n}(x,y) = \begin{cases} H_n(y) & si\ x \ge y, \\ \sum_{l=1}^{n} C_n^l\ F(x)^l\big(F(y)-F(x)\big)^{n-l} & si\ x < y, \end{cases}$$

$$= \begin{cases} H_n(y) & si\ x \ge y, \\ H_n(y) - \big(F(y)-F(x)\big)^n & si\ x < y, \end{cases}$$

$$h_{1,n}(x,y) = n\,(n-1)\big(F(y)-F(x)\big)^{n-2} f(x)f(y)\mathbb{1}_{\{x<y\}}.$$

Preuve :

On a $H_{i,j}(x,y) = P\big(X_{(i)} \le x, X_{(j)} \le y\big)$.

Si $x \ge y$ alors $\big\{X_{(j)} \le y\big\} \subset \big\{X_{(i)} \le x\big\}$ et donc $H_{i,j}(x,y) = H_j(y)$.

Si $x < y$ alors l'événement $A = \big\{X_{(i)} \le x\ ,\ X_{(j)} \le y\big\}$ peut s'écrire sous la forme :

$$A = \cup_{k=j}^{n} \cup_{l=i}^{k} A_{l,k}$$

où $A_{l,k}$ désigne l'événement: l des n variables X_m sont inférieures ou égales à x, $(k-l)$ sont supérieures à x et inférieures ou égales à y, les $(n-k)$ autres sont supérieures à y.

D'où :

$$P(A) = \sum_{k=j}^{n}\sum_{l=i}^{k} P(A_{l,k})$$

$$= \sum_{k=j}^{n}\sum_{l=i}^{k} C_n^l C_{n-l}^{k-l}\big(F(x)\big)^l\big(F(y)-F(x)\big)^{k-l}\big(1-F(y)\big)^{n-k}$$

$$= \sum_{k=j}^{n}\sum_{l=i}^{k} C_n^k C_k^l\big(F(x)\big)^l\big(F(y)-F(x)\big)^{k-l}\big(1-F(y)\big)^{n-k}$$

$$= \sum_{k=j}^{n}\sum_{l=i}^{k} \frac{n!}{l!\,(k-l)!\,(n-k)!}\big(F(x)\big)^l\big(F(y)-F(x)\big)^{k-l}\big(1-F(y)\big)^{n-k}.$$

Finalement

$$H_{i,j}(x,y) = \begin{cases} H_j(y) & si\ x \geq y, \\ \sum_{k=j}^{n}\sum_{l=i}^{k} \frac{n!}{l!\,(k-l)!\,(n-k)!}\big(F(x)\big)^l\big(F(y)-F(x)\big)^{k-l}\big(1-F(y)\big)^{n-k} & si\ x < y. \end{cases}$$

On obtient la densité conjointe (dans le cas où F admet une densité f) par la formule

$$h_{i,j}(x,y) = \frac{\partial^2}{\partial x \partial y} H_{i,j}(x,y).$$

Grâce aux éliminations de termes, on obtient le résultat du théorème.

<u>Remarque</u>

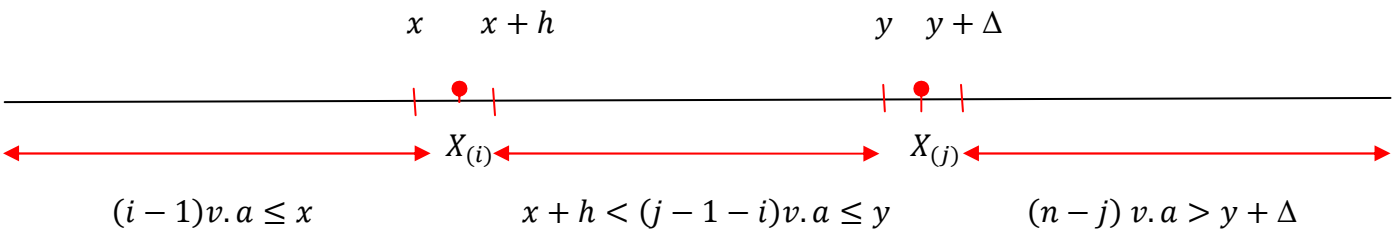
On a :

$$h_{i,j}(x,y) = \lim_{\substack{h\to 0\\ \Delta\to 0}} \frac{H_{i,j}(x+h,y+\Delta) - H_{i,j}(x,y+\Delta) - H_{i,j}(x+h,y) + H_{i,j}(x,y)}{h\Delta}$$

$$= \lim_{\substack{h\to 0\\ \Delta\to 0}} \frac{P\big(x < X_{(i)} \leq x+h,\ y < X_{(j)} \leq y+\Delta\big)}{h\Delta}.$$

Si $x \geq y$ alors $h_{i,j}(x,y) = 0$.

Si $x \leq y$ alors



$$P\big(x < X_{(i)} \leq x+h, y < X_{(j)} \leq y+\Delta\big) =$$

$$= C_n^{i-1}C_{n-i+1}^1 C_{n-i}^{j-i-1}C_{n-j+1}^1 (F(x))^{i-1}\big(F(x+h)-F(x)\big)\big(F(y)-F(x+h)\big)^{j-i-1}\big(F(y+\Delta)-F(y)\big)\big(1-F(y+\Delta)\big)^{n-j}.$$

3.5 Généralisation:

Soit le vecteur aléatoire $(X_{(i_1)}, \dots, X_{(i_k)})$ de densité conjointe $h_{i_1,\dots,i_k}$, $1 \le i_1 < \cdots < i_k \le n,\ k = \overline{1,n}$

Théorème:

$h_{i_1,\dots,i_k}(x_1, \dots, x_k)$

$$= \frac{n!\ \ (F(x_1))^{i_1-1}}{(i_1-1)!\,(n-i_k)!\prod_{j=1}^{k-1}(i_{j+1}-i_j-1)!}\prod_{j=1}^{k-1}\big(F(x_{j+1})-F(x_j)\big)^{i_{j+1}-i_j-1}\,(1-F(x_k))^{n-i_k}\prod_{j=1}^{k} f(x_j)\ \mathbb{1}_{\{x_{i_1}<x_{i_2}<\dots<x_{i_k}\}}.$$

En particulier la densité conjointe de l'échantillon ordonnée $(X_{(1)}, \dots, X_{(n)})$ est donnée par :

$$h_{1,\dots,n}(x_1, \dots, x_n) = n!\prod_{j=1}^{n} f(x_j)\,\mathbb{1}_{\{x_1<x_2<\dots<x_n\}}.$$

3.6 Distribution de l'étendue :

Soient G et g respectivement la fonction de répartition et la densité de l'étendue W définie par :

$$W = X_{(n)} - X_{(1)}.$$

Théorème :

$\forall w \ge 0$

$$G(w) = n\int_{-\infty}^{+\infty} f(u)\big(F(u+w)-F(u)\big)^{n-1}du,$$

$$g(w) = n(n-1)\int_{-\infty}^{+\infty} f(u)f(u+v)\big(F(u+w)-F(u)\big)^{n-2}du.$$

Preuve :

Soit $h_{1,n}$ la densité conjointe de $(X_{(1)}, X_{(n)})$. On considère le changement de variables

$$\varphi(X_{(1)}, X_{(n)}) = (U, W) = (X_{(1)}, X_{(n)} - X_{(1)})$$

$$\varphi: \mathbb{R}^2 \to \mathbb{R}\times\mathbb{R}^+$$

$$(x, y) \to (u, w) = (x, y-x).$$

D'après le théorème de changement de variables, on a

$$f_{U,W}(u,w) = f_{X_{(1)},X_{(n)}}\big(\varphi^{-1}(u,w)\big)\big|J_{\varphi^{-1}}\big|$$

avec $\varphi^{-1}(u,w) = (u, u+w) = (x,y)$ et $J_{\varphi^{-1}} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial w} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1.$

$$f_{U,W}(u,w) = h_{1,n}(u, u+w) = n(n-1)\big(F(u+v)-F(u)\big)^{n-2}f(u)f(u+w)\mathbb{1}_{\{w>0\}}.$$

On obtient la densité de W par la relation $g(w) = f_W(w) = \int_{-\infty}^{+\infty} f_{U,W}(u,w)\,du$.

Si $w \leq 0$, alors $g(w) = 0$.

Si $w > 0$, $g(w) = n(n-1)\int_{-\infty}^{+\infty} f(u)f(u+w)\big(F(u+w) - F(u)\big)^{n-2}du$.

Pour obtenir la fonction de répartition de W, on applique la relation $G(w) = \int_{-\infty}^{w} g(x)dx$.

Si $w \leq 0$, alors $G(w) = 0$.

Si $w > 0, G(w) = \int_0^w n(n-1)\int_{-\infty}^{+\infty} f(u)f(x+u)\big(F(x+u) - F(u)\big)^{n-2}du\,dx$

$$= n\int_{-\infty}^{+\infty} f(u)\int_0^w (n-1)f(x+u)\big(F(x+u) - F(u)\big)^{n-2}\,dxdu$$

$$= n\int_{-\infty}^{+\infty} f(u)\big(F(w+u) - F(u)\big)^{n-1}du.$$

Remarque :

Dans le cas où $X \rightsquigarrow N(m, \sigma^2)$, la loi (fonction de répartition) de W est tabulée.

3.7 Généralisation : Les (i, j)-étendues

Soient $\big(X_{(1)}, \dots, X_{(n)}\big)$ l'échantillon ordonné de $(X_1, \dots, X_n)$ et i, $j \in \{1,2, \dots, n\}$ tels que $i < j$.
On appelle (i, j)-étendue la statistique $W_{i,j}$ définie par :

$$W_{i,j} = X_{(j)} - X_{(i)}.$$

Pour $i = 1, j = n$, on obtient l'étendue $W = W_{1,n}$.

La fonction de répartition $G_{i,j}$ et la dentsité $g_{i,j}$ de $W_{i,j}$ sont obtenues de la même manière que celles de W :

$$g_{i,j}(w) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}\int_{-\infty}^{+\infty}\big(F(u)\big)^{i-1}\big(F(u+w) - F(u)\big)^{j-i-1}\big(1 - F(u+w)\big)^{n-j}f(u)f(u+w)du.$$

Théorème :

Si $(X_1, \dots, X_n)$ est un n-échantillon de loi $U_{[0,1]}$ alors les différences $W_{i,i+1} = X_{(i+1)} - X_{(i)}$ sont de même loi que $X_{(1)}(\rightsquigarrow \mathcal{B}_{[0,1]}(1, n))$.

Théorème :

si $(X_1, \dots, X_n)$ est un n-échantillon de loi $\xi(\lambda)$ alors les $W_{i,i+1}$ sont des variables aléatoires exponentielles indépendantes telles que

$$W_{i,i+1} \rightsquigarrow \xi\big(\lambda(n-i)\big).$$

Remarque :

Dans le cas gaussien :

1. Les $E\big(X_{(i)}\big), Var\big(X_{(i)}\big),\ Cov\big(X_{(i)}, X_{(j)}\big), i < j,\ et\ E(W)$ sont tabulées,
2. Les $E\big(X_{(i)}\big)$ sont appelées des scores normaux.

3.8 Loi des composées des statistiques d'ordre et de la fonction de répartition

Soient

- $Z_i = F\left(X_{(i)}\right), i = \overline{1,n}$,
- $l_{i_1,\cdots,i_k}$ la densité conjointe de $\left(Z_{i_1}, \dots, Z_{i_k}\right)$, $k = \overline{1,n}$.

Théorème :

1. $F(X) \rightsquigarrow U_{[0,1]}$.
2. $Z_i \rightsquigarrow \mathcal{B}_{[0,1]}(i, n-i+1)$
3.

$l_{i_1,\cdots,i_k}(z_1, \dots, z_k)$

$$= \frac{n!\ \ (z_1)^{i_1-1}}{(i_1-1)!\,(n-i_k)!\prod_{j=1}^{k-1}\left(i_{j+1}-i_j-1\right)!}\prod_{j=1}^{k-1}\left(z_{j+1}-z_j\right)^{i_{j+1}-i_j-1}(1-z_k)^{n-i_k}\ \mathbb{I}_{\{0\le z_1<z_2<\dots<z_k\le 1\}}.$$

Preuve

Les variables $X_1, \dots, X_n$ étant i.i.d. de fonction de répartition F alors $U_1 = F(X_1), \dots, U_n = F(X_n)$ sont i.i.d de loi $U_{[0,1]}$. La fonction F est croissante donc $U_{(1)} = Z_1 = F\left(X_{(1)}\right), \dots, U_{(n)} = Z_n = F\left(X_{(n)}\right)$ sont les statistiques d'ordre associées à un échantillon de la loi $U_{[0,1]}$. Par suite, on applique les résultats établis pour les statistiques d'ordre.

4 Statistiques de rang :

En statistique non paramétrique, les valeurs observées sont transformées en une série de rangs.

4.1 Définition :

Soient $(X_1, \dots, X_n)$ un n-échantillon d'une v.a. X de f.r. F et $\left(X_{(1)}, \dots, X_{(n)}\right)$ l'échantillon ordonné. Soit R_i le rang de X_i dans l'échantillon ordonné, $i = \overline{1,n}$. Il est donné par

$$R_i = 1 + \sum_{j=1}^{n} \mathbb{1}_{\{X_j < X_i\}}.$$

Le vecteur des rangs $(R_1, \dots, R_n)$ est une permutation de $\{1, \dots, n\}$. S'il ya des ex aequos, le vecteur des rangs n'est pas unique. On peut attribuer par exemple aux ex aequos un rang moyen.

Exemple :

$(x_1, \dots, x_7)$ est une réalisation de $(X_1, \dots, X_7)$, $\left(x_{(1)}, \dots, x_{(7)}\right)$ est l'échantillon ordonné et $(r_1, \dots, r_7)$ est une réalisation de $(R_1, \dots, R_7)$.

x_i	1.1	2.3	1.4	1.7	0.9	1.8	2.1
$x_{(i)}$	0.9	1.1	1.4	1.7	1.8	2.1	2.3
r_i	2	7	3	4	1	5	6

Si on remplace x_4 par 1.4 alors on obtient

x_i	1.1	2.3	1.4	1.4	0.9	1.8	2.1
$x_{(i)}$	0.9	1.1	1.4	1.4	1.8	2.1	2.3
r_i	2	7	3.5	3.5	1	5	6

4.2 Loi de R_i et $(R_1, \dots, R_n)$:

Soit $(R_1, \dots, R_n)$ le vecteur des rangs associé à l'échantillon $(X_1, \dots, X_n)$, R_i étant le rang de X_i dans l'échantillon ordonné $(X_{(1)}, \dots, X_{(n)})$, $i = \overline{1,n}$.

Le vecteur des rangs $(R_1, \dots, R_n)$ est à valeurs dans $\sum_n$ l'ensemble des permutations de $\{1, \dots, n\}$.

Remarque :

Si la réalisation de l'échantillon ne présente pas d'ex aequos, les entiers $r_1, \dots, r_n$ correspondants sont distincts et il existe une et une seule permutation σ de l'ensemble $\sum_n$ des permutations de $\{1, \dots, n\}$ telle que

$$r_1 = \sigma(1), \dots, r_n = \sigma(n) \text{ et } 1 = \sigma^{-1}(r_1), \dots, n = \sigma^{-1}(r_n).$$

Théorème :

Soit $(X_1, \dots, X_n)$ un n-échantillon d'une v.a. absolument continue.

1. La loi de $(R_1, \dots, R_n)$ est uniforme sur $\sum_n$.
2. Pour tous $i, j, k, l \in \{1, \dots, n\}, i \neq j, k \neq l$, $P(R_i = k, R_j = l) = \frac{1}{n(n-1)}$, $Cov(R_i, R_j) = -\frac{n+1}{12}$.
3. Pour tous $i_1, \dots, i_k, l_1, \dots, l_k \in \{1, \dots, n\}, i_j \neq i_m, l_j \neq l_m, j \neq m, P(R_{i_1} = l_1, \dots, R_{i_k} = l_k) = \frac{(n-k)!}{n!}$.
4. Les variables R_i sont uniformes sur $\{1, \dots, n\}$, $E(R_i) = \frac{n+1}{2}, Var(R_i) = \frac{n^2-1}{12}$.
5. Les vecteurs $(X_{(1)}, \dots, X_{(n)})$ et $(R_1, \dots, R_n)$ sont indépendants.

Preuve :

1. Soit $\sigma \in \sum_n$, $\sigma = (\sigma(1), \dots, \sigma(n))$. L'événement $[(R_1, \dots, R_n) = (\sigma(1), \dots, \sigma(n))]$ est équivalent à $[X_{(1)} = X_{\sigma^{-1}(1)}, \dots, X_{(n)} = X_{\sigma^{-1}(n)}]$. D'où

   $$P[(R_1, \dots, R_n) = (\sigma(1), \dots, \sigma(n))] = P[X_{\sigma^{-1}(1)} < \dots < X_{\sigma^{-1}(n)}].$$

   Puisque les v.a. $X_1, \dots, X_n$ sont échangeables, alors $\forall \sigma \in \sum_n$,

   $$P[X_{\sigma^{-1}(1)} < \dots < X_{\sigma^{-1}(n)}] = P[X_1 < \dots < X_n] \text{ et donc } P[(R_1, \dots, R_n) = (\sigma(1), \dots, \sigma(n))] = \frac{1}{n!}.$$

2. Pour tous $i, j, k, l \in \{1, \dots, n\}, i \neq j, k \neq l$,

   $$P(R_i = k, R_j = l) = \sum_{\sigma \in \sum_n | \sigma(i) = k, \sigma(j) = l} P[(R_1, \dots, R_n) = (\sigma(1), \dots, \sigma(n))] = \frac{(n-2)!}{n}.$$

3. Pour tout $k \in \{1, \dots, n\}$,

$$P(R_i = k) = \sum_{\sigma \in \sum_n | \sigma(i)=k} P\big[(R_1, \dots, R_n) = \big(\sigma(1), \dots, \sigma(n)\big)\big] = \frac{(n-1)!}{n} = \frac{1}{n}.$$

5 Tests statistiques

5.1 Principe d'un test d'hypothèses :

Un test d'hypothèse consiste à établir une règle de décision pour choisir entre deux hypothèses (scénarios), H_0 et H_1. L'hypothèse H_0, appelée hypothèse nulle, est la plus plausible (à priori vraie). L'hypothèse H_1, appelée hypothèse alternative, est l'hypothèse que l'on veut démontrer. Choisir d'accepter ou de rejeter H_0 peut mener à commettre deux types d'erreur :

- Erreur de 1ère espèce « $\boldsymbol{\alpha}$ » qui est la probabilité de rejeter H_0 alors qu'elle est vraie,
- Erreur de 2ème espèce « $\boldsymbol{\beta}$ » qui est la probabilité d'accepter H_0 alors qu'elle est fausse.

Réalité / **Décision**	$\boldsymbol{H_0}$ **Vraie**	$\boldsymbol{H_1}$ **Vraie**
$\boldsymbol{H_0}$ **Acceptée**	« Correct » $\boldsymbol{1-\alpha} = P$(accepter H_0/ H_0vraie)	« Manque de puissance » $\boldsymbol{\beta} = P$(accepter H_0/ H_0fausse) Risque de seconde espèce β
$\boldsymbol{H_1}$ **Acceptée**	« Rejet à tort » $\boldsymbol{\alpha} = P$(rejeter H_0/ H_0vraie) Risque de première espèce α	« Puissance du Test » $1-\beta$ $\boldsymbol{1-\beta} = P$(rejeter H_0/ H_0fausse)

Résoudre un problème de test revient à déterminer sa région critique, région de rejet de H_0, qui est basée sur une statistique dont on connait la loi sous H_0. En pratique, l'erreur de première espèce α est fixée au préalable. Généralement, on prend $\alpha = 0.1,\ 0.05,\ 0.01$.

5.2 Seuil et $p-valeur$:

Le seuil est la probabilité α, fixée à priori, que le test rejette H_0 à tort,

$$\boldsymbol{\alpha} = P_{H_0}(\text{rejeter } H_0) = P(\text{rejeter } H_0/\ H_0\text{vraie}).$$

La valeur prise par la statistique de test est calculée sur la base de données recueillies et la réponse sera binaire : rejet ou non de H_0. On préfère souvent calculer le seuil limite auquel H_0 aurait été rejetée compte tenu de la valeur de la statistique de test.

Définition :

Soient H_0 l'hypothèse nulle, T la statistique de test et F_0 sa fonction de répartition sous H_0. On suppose que F_0 est continue. Selon l'hypothèse alternative H_1 le test est bilatéral ou unilatéral.

La région critique W et la p-valeur d'une valeur t prise par T, notée $p(t)$, sont données respectivement par

1. Pour un test bilatéral (rejet des valeurs trop écartées)

$$W: |T| > k_\alpha \text{ telle que } P_{H_0}(|T| > k_\alpha) = \alpha,$$

$$\alpha_0 = p(t) = \begin{cases} 2\,F_0(t) & si\ F_0(t) < 0.5, \\ 2(1 - F_0(t)) & si\ F_0(t) \geq 0.5, \end{cases}$$

2. Pour un test bilatéral à droite (rejet des valeurs trop grandes)

$$W: T > k_\alpha \text{ telle que } P_{H_0}(T > k_\alpha) = \alpha,$$

$$\alpha_0 = p(t) = P_{H_0}(\mathrm{T} > t) = 1 - F_0(t),$$

3. Pour un test bilatéral à gauche (rejet des valeurs trop petites)

$$W: T < k_\alpha \text{ telle que } P_{H_0}(< k_\alpha) = \alpha$$

$$\alpha_0 = p(t) = P_{H_0}(\mathrm{T} < t) = P_{H_0}(\mathrm{T} \leq \mathrm{t}) = F_0(t) \text{ (continuité de } F_0).$$

En pratique, pour une statistique de test de fonction de répartition F_0 sous H_0, on définira souvent la p-valeur d'une valeur t par $\alpha_0 = p(t) = \min(F_0(t), 1 - F_0(t))$.

La connaissance de la p-valeur rend inutile le calcul préalable de la région critique. En effet, si α_0 est la p-valeur d'une observation t sous H_0, on obtient un test de seuil α par la règle de rejet:

$$\text{Rejeter } H_0 \Leftrightarrow \alpha_0 \leq \alpha.$$

Remarque:

Dans le cas d'une statistique de test discrète, il faut inclure la valeur observée dans l'intervalle dont on calcule la probabilité :

- Pour un test unilatéral à droite : $\alpha_0 = P_{H_0}(\mathrm{T} \geq \mathrm{t})$,
- Pour un test unilatéral à gauche : $\alpha_0 = P_{H_0}(\mathrm{T} \leq \mathrm{t})$.