Tests non paramétriques sur les médianes

Tests statistiques

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

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 $\alpha = 0.1,\ 0.05,\ 0.01$.

Seuil et $p-valeur$:

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

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

Notons que le seuil α est toujours réalisable dans le cas continu mais ne l'est pas toujours dans le cas discret.

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)

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

$$= 2\min\big(F_0(t), 1 - F_0(t)\big)$$

2. Pour un test unilatéral 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 unilatéral à gauche (rejet des valeurs trop petites)

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

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

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 : $\alpha_0 = P_{H_0}(T \geq t)$,
- Pour un test unilatéral à gauche : $\alpha_0 = P_{H_0}(T \leq t)$,
- Pour un test bilatéral : $\alpha_0 = 2 \min\left(P_{H_0}(T \geq t), P_{H_0}(T \leq t)\right)$.

Test du signe

Soient $(X_1, \dots, X_n)$ un n-échantillon d'une variable aléatoire X absolument continue, de médiane inconnue M (supposée unique) et M_0 une valeur donnée. On veut tester :

$$H_0: M = M_0 \quad contre \quad H_1: M \neq M_0 \ (H_1: M > M_0 \text{ ou } H_1: M < M_0).$$

On associe un signe $(+)$ aux valeurs supérieures à M_0 et un signe $(-)$ aux valeurs inférieures à M_0 et on note T^+ le nombre de $(+)$ et T^- le nombre de $(-)$

$$T^+ = \sum_{i=1}^{n} \mathbb{1}_{\{X_i > M_0\}}, \quad T^- = \sum_{i=1}^{n} \mathbb{1}_{\{X_i < M_0\}}.$$

On suppose dans un premier temps que $X_i \neq M_0, \forall\, i = \overline{1,n}$ (ce qui permet d'écrire $T^+ + T^- = n$).

Sous H_0, $P\{X_i > M_0\} = P\{X_i < M_0\} = \frac{1}{2}, \forall\, i = \overline{1,n}$ et $T^+,\ T^- \rightsquigarrow \mathcal{B}\left(n, \frac{1}{2}\right)$.

Si H_1 est de la forme $M > M_0$ alors une trop forte valeur de T^+ (trop faible valeur de T^-) incitera au rejet de H_0 en faveur de H_1. Dans ce cas la région critique est de la forme $\{T^+ \geq t_\alpha^*\}$ $(ou\ T^- \leq t_\alpha^{**})$, t_α^* étant le plus petit entier tel que :

$$P_{H_0}(T^+ \geq t_\alpha^*) = \sum_{j=t_\alpha^*}^{n} C_n^j (0.5)^n \leq \alpha.$$

Si H_1 est de la forme $M < M_0$ alors une trop faible valeur de T^+ (trop forte valeur de T^-) incitera au rejet de H_0. La région de rejet est de la forme $\{T^+ \leq t_\alpha^*\}$ $(ou\ T^- \geq t_\alpha^{**})$, t_α^* étant le plus grand entier positif tel que

$$P_{H_0}(T^+ \leq t_\alpha^*) = \sum_{j=0}^{t_\alpha^*} C_n^j (0.5)^n \leq \alpha.$$

Si H_1 est de la forme $M \neq M_0$, H_0 est rejetée lorsque $T^+ \notin \left]t_{\alpha/2}^*, n - t_{\alpha/2}^*\right[$, avec $t_{\alpha/2}^*$ est tel que

$$P_{H_0}\left(T^+ \leq t_{\alpha/2}^*\right) \leq \alpha/2.$$

En effet, dans le cas d'un test bilatéral, H_0 est rejetée pour une trop faible valeur de T^- ou une trop faible valeur de T^+ ($T^+ \leq c_1$ ou $T^- \leq c_2$). Comme $T^- = n - T^+$ alors $(T^- \leq c_2) \Leftrightarrow (T^+ \geq n - c_2)$. Pour un seuil α fixé, c_1 et c_2 satisfont $P_{H_0}(T^+ \leq c_1) + P_{H_0}(T^- \leq c_2) \leq \alpha$, et du fait que T^+ et T^- sont de même loi , on peut poser $P_{H_0}(T^+ \leq c_1) = P_{H_0}(T^- \leq c_2)$ et par suite $P_{H_0}(T^+ \leq c_1) \leq \alpha/2$ et $c_1 = c_2 = t^*_{\alpha/2}$.

Il existe une table de valeurs critiques pour le test de signe.

Soient T^+, T^- le nombre de $(+)$ et de $(-)$ respectivement.

On pose

$$T = \begin{cases} Min(T^+, T^-) & si\ H_1 : M \neq M_0 \\ T^- & si\ H_1 : M > M_0 \\ T^+ & si\ H_1 : M < M_0. \end{cases}$$

Règle de décision : Si $T \leq t^*_\alpha$ alors rejeter H_0 sinon accepter H_0.

La table des valeurs critiques pour le test de signe donne la valeur t^*_α telle que

$$P_{H_0}(T \leq t^*_\alpha) \leq \alpha.$$

Le niveau de signification d'un test bilatéral est égal à 2 fois le niveau de signification d'un test unilatéral. De plus la valeur critique t^*_α d'un test bilatéral de niveau de signification α est égale à la valeur critique $t^*_{\alpha/2}$ d'un test unilatéral à gauche de niveau de signification $\alpha/2$.

La p −valeur d'une valeur t prise par T^+ est :

$$\alpha_0 = \begin{cases} P_{H_0}(T^+ \geq t) = 1 - F_{\mathcal{B}\left(n,\frac{1}{2}\right)}(t-1) & si\ \ H_1 : M > M_0, \\ P_{H_0}(T^+ \leq t) = F_{\mathcal{B}\left(n,\frac{1}{2}\right)}(t) & si\ H_1 : M < M_0, \\ 2Min(P_{H_0}(T^+ \leq t), P_{H_0}(T^+ \geq t)) & si\ \ H_1 : M \neq M_0. \end{cases}$$

Notons que dans le cas d'un test bilatéral, on a

$$\alpha_0 = \begin{cases} 2P_{H_0}(T^+ \leq t) & si \quad F_{\mathcal{B}\left(n,\frac{1}{2}\right)}(t) > 0.5 \quad \left(t < \frac{n}{2}\right), \\ 2P_{H_0}(T^+ \geq t) & si \quad F_{\mathcal{B}\left(n,\frac{1}{2}\right)}(t) \geq 0.5 \quad \left(t > \frac{n}{2}\right). \end{cases}$$

Si $\alpha_0 \leq \alpha$ ⟶ rejeter H_0 sinon accepter H_0.

Exemple :

On dispose des observations suivantes issues d'une variable aléatoire X continue de médiane inconnue M :

x_i	0.28	0.18	0.24	0.30	0.4	0.36	0.15	0.42	0.23	0.48

Tester au seuil $\alpha = 5\%$

i) $H_0: M = 0.25 \ vs \ H_1: M \neq 0.25$

Signe$(x_i - 0.25)$	+	-	-	+	+	+	-	+	-	+

On a $T^+_{obs} = t^+ = 6 > \frac{n}{2}$, $T^-_{obs} = t^- = 4$ et $sous \quad H_0$, $T^+ \rightsquigarrow \mathcal{B}\left(10, \frac{1}{2}\right)$.

- Avec la $p-$valeur

$$\alpha_0 = 2Min(P_{H_0}(T^+ \geq 6), P_{H_0}(T^+ \leq 6)) = 2Min(1 - F_{\mathcal{B}\left(10,\frac{1}{2}\right)}(5), F_{\mathcal{B}\left(10,\frac{1}{2}\right)}(6))$$

$$= 2Min(1 - 0.6230, 0.8281) = 0.754 > 0.05 \longrightarrow \text{accepter } H_0.$$

- Avec la valeur critique

$T_{obs} = t = min(t^+, t^-) = t^- = 4.$

De la table, on lit $t^*_{0.05} = 1$ pour $n = 10$.

Comme $t > t^*_{0.05}$ alors on accepte H_0.

ii) $H_0: M = 0.25 \ vs \ H_1: M > 0.25$

- Avec la $p-$valeur

$\alpha_0 = P_{H_0}(T^+ \geq 6) = 1 - F_{\mathcal{B}\left(10,\frac{1}{2}\right)}(5) = 0.377 > 0.05 \longrightarrow$ accepter H_0.

- Avec la valeur critique

  On a $t = t^- = 4$ et $t^*_{0.05} = 1$ pour $n = 10$. Comme $t > t^*_{0.05}$ alors on accepte H_0.

iii) $H_0: M = 0.25 \ vs \ H_1: M < 0.25$

- Avec la $p-$valeur

  $\alpha_0 = P_{H_0}(T^+ \leq 6) = F_{\mathcal{B}\left(10,\frac{1}{2}\right)}(6) = 0.8281 > 0.05 \longrightarrow$ accepter H_0.

- Avec la valeur critique

  On a $t = t^+ = 4$ et $t^*_{0.05} = 1$ pour $n = 10$. Comme $t > t^*_{0.05}$ alors on accepte H_0.

Remarque :

Il est possible que certaines valeurs de l'échantillon soient égales à M_0 . Plusieurs stratégies sont possibles, entre autres ignorer (éliminer) ces valeurs et appliquer le test sur les données restantes.

Exemple :

x_i	0.28	0.18	0.24	0.30	0.4	0.36	0.15	0.42	0.23	0.48

Tester, au seuil $\alpha = 0.05$, $H_0: M = 0.28 \ \ v.s \ H_1: M \neq 0.28$

Signe$(x_i - 0.28)$	/	-	-	+	+	+	-	+	-	+

On a $t^+ = 5$, $t^- = 4$ et $sous \ H_0$, $T^+ \rightsquigarrow \mathcal{B}\left(9, \frac{1}{2}\right)$.

- Avec la p-valeur

  $\alpha_0 = 2P_{H_0}(T^+ \geq 5) = 2\left(1 - F_{\mathcal{B}\left(9,\frac{1}{2}\right)}(4)\right) = 1 > 0.05$, donc on accepte H_0.

- Avec la valeur critique

  $t = \min(t^+, \ t^-) = 4$ et $t^*_{0.05} = 1 < t$, donc on accepte H_0.

Exemple :

Les tailles (en cm) de 15 étudiants sélectionnés au hasard dans une école préparatoire sont données dans le tableau suivant :

x_i	132	134	138	139	142	132	140	136	135	140	139	132	131	136	138

Soit M la taille médiane des étudiants de l'école préparatoire. Tester au niveau $\alpha = 5\%$

i) $H_0: M = 135 \; vs \; H_1: M \neq 135$

Signe$(x_i - 135)$	-	-	+	+	+	-	+	+	/	+	+	-	-	+	+

On a $t^+ = 9, \; et \; sous \; H_0, \; T^+ \rightsquigarrow \mathcal{B}\left(14, \frac{1}{2}\right)$. Donc

$$\alpha_0 = 2P_{H_0}(T^+ \geq 9) = 2\sum_{k=9}^{14} C_{14}^k (0.5)^{14} = 0.4239502 > 0.05 \rightarrow accepter \; H_0.$$

ii) $H_0: M = 135 \; vs \; H_1: M < 135$

$$\alpha_0 = F_{\mathcal{B}\left(14,\frac{1}{2}\right)}(9) = 0.9102173 > 0.05 \rightarrow accepter \; H_0.$$

iii) $H_0: M = 137 \; vs \; H_1: M > 137$

Signe$(x_i - 137)$	-	-	+	+	+	-	+	-	-	+	+	-	-	-	+

$$sous \; H_0, \qquad T^+ \rightsquigarrow \mathcal{B}\left(15, \frac{1}{2}\right)$$

$t^+ = 7, \; \alpha_0 = 1 - P_{H_0}(T^+ \leq 6) = P_{H_0}(T^+ \geq 7) = 0.696 > 0.05 \rightarrow accepter \; H_0.$

iv) $H_0: M = 141 \; vs \; H_1: M < 141$

Signe$(x_i - 141)$	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-

$t^+ = 1,\ \ \alpha_0 = F_{\mathcal{B}\left(15,\frac{1}{2}\right)}(1) = 0.0014 < 0.05\ on\ accepte\ H_1.$

Pour les valeurs critiques, on a :

- $H_0: M = 135\ vs\ \ H_1: M \neq 135$

  $t^+ = 9\ et\ \ t^- = 5$ donc $t = min(t^+, t^-) = min(5,9) = 5.$

De la table on lit pour $n = 14$, $t^*_{0.05} = 2.$

$$t = 5 > 2 \rightarrow on\ accepte\ H_0$$

- $H_0: M = 137\ vs\ \ H_1: M > 137$

$t^+ = 7,\ \ t^- = 8,\ \ t = t^- = 8,\ \ et\ \ t^*_{0.05} = 3\ \ pour\ n = 15.$

Comme $t > 3\ alors\ on\ accepte\ H_0.$

- $H_0: M = 141\ vs\ \ H_1: M < 141$

  $t = t^+ = 1,\ \ t^*_{0.05} = 3, n = 15,\ \ t^+ < 3 \rightarrow on\ rejette\ H_0.$

<u>Approximation normale :</u>

Sous $H_0: M = M_0$ et pour les grands échantillons $(n \geq 30)$ on peut utiliser la statistique

$$Z = \frac{T^+ - \mathrm{E}(T^+)}{\sqrt{\mathrm{var}(T^+)}} = \frac{T^+ - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

qui suit une loi normale centrée réduite.

La p-valeur α_0d'une valeur *t* observée de T^+ est telle que

$$\alpha_0 = \begin{cases} 2Min\left(\left(1-\Phi\left(\dfrac{t-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)\right), \Phi\left(\dfrac{t-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)\right) & si\ H_1: M \neq M_0, \\ 1-\Phi\left(\dfrac{t-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) & si\ H_1: M > M_0, \\ \Phi\left(\dfrac{t-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) & si\ H_1: M < M_0. \end{cases}$$

Les régions critiques associées sont respectivement $\begin{cases} |Z| > z_{1-\alpha/2}, & si\ H_1: M \neq M_0 \\ Z > z_{1-\alpha}, & si\ H_1: M > M_0, \\ Z < z_{\alpha}. & si\ H_1: M < M_0. \end{cases}$

z_α : quantile d'ordre α de la loi $N(0,1)$ avec $z_\alpha = -z_{1-\alpha}$.

Une correction de continuité peut s'avérer nécessaire dans ce cas

<u>Correction de continuité :</u>

T^+ est une variable aléatoire discrète alors pour toute valeur $u \in [0,1[$, on a

$$P(a \leq T^+ \leq b) = P(a-u \leq T^+ \leq b+u).$$

En particulier si la région de rejet est de la forme $\{T^+ \geq c\}$ (dans le cas $H_1: M > M_0$) alors pour toute valeur $u \in [0,1[$, on a $P_{H_0}(T^+ \geq c) = P_{H_0}(T^+ \geq c-u)$.

De la même façon si la région de rejet est de la forme $\{T^+ \leq c\}$ ($dans\ le\ cas\ H_1: M < M_0$) alors pour toute valeur $u \in [0,1[$, on a $P_{H_0}(T^+ \leq c) = P_{H_0}(T^+ \leq c+u)$.

La correction de continuité consiste à remplacer par défaut la valeur de u par $\frac{1}{2}$.

Dans le cas $\{T^+ \geq c\}$, on cherche c tel que

$$P_{H_0}(T^+ \geq c - 0.5) = \alpha \Leftrightarrow P_{H_0}\left(\frac{T^+ - \frac{n}{2}}{\frac{\sqrt{n}}{2}} \geq \frac{c - 0.5 - \frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) = \alpha,$$

$$\Leftrightarrow 1 - \Phi\left(\frac{c - 0.5 - \frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) = \alpha$$

$$c = \frac{\sqrt{n}}{2} z_{1-\alpha} + 0.5 + \frac{n}{2}.$$

Dans le cas $\{ T^+ \leq c\}$, on cherche c tel que

$$P_{H_0}(T^+ \leq c + 0.5) = \alpha \Leftrightarrow P_{H_0}\left(\frac{T^+ - \frac{n}{2}}{\frac{\sqrt{n}}{2}} \leq \frac{c + 0.5 - \frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) = \alpha$$

$$\Leftrightarrow \Phi\left(\frac{c + 0.5 - \frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) = \alpha$$

$$c = \frac{\sqrt{n}}{2} z_{\alpha} - 0.5 + \frac{n}{2}.$$

Dans le cas d'un test bilatéral, on obtient

$$T^+ \leq \frac{\sqrt{n}}{2} z_{\alpha/2} - 0.5 + \frac{n}{2} \text{ ou } T^+ \geq \frac{\sqrt{n}}{2} z_{1-\alpha/2} + 0.5 + \frac{n}{2}.$$

La p-valeur α_0d'une valeur t observée de T^+ est telle que

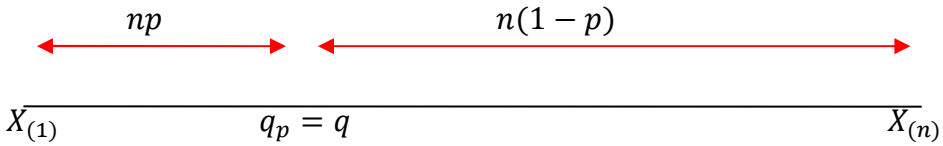
$$\alpha_0 = \begin{cases} 2Min\left(\left(1-\Phi\left(\dfrac{t-0.5-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)\right), \Phi\left(\dfrac{t+0.5-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)\right) & si\ H_1: M \neq M_0, \\ 1-\Phi\left(\dfrac{t-0.5-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) & si\ H_1: M > M_0, \\ \Phi\left(\dfrac{t+0.5-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right) & si\ H_1: M < M_0. \end{cases}$$

Généralisation : Test de signe pour un quantile

Soient q_p le quantile d'ordre p de l'échantillon $(X_1, \dots, X_n)$ et q une valeur fixée. On peut généraliser le test de signe au problème

$$H_0: q_p = q \ vs\ \ H_1: q_p \neq q \quad \left(q_p > q \ \ ou\ q_p < q\right)$$

Il suffit d'assigner un signe $(-)$ aux valeurs inférieures à q, un signe $(+)$ aux valeurs supérieures à q et de considérer le nombre de $(+)$ T^+ qui suit sous H_0 une loi binomiale $\mathcal{B}(n, 1-p)$



(On peut également considérer T^- le nombre de (-) qui suit sous H_0 une loi binomiale $\mathcal{B}(n, p)$)

Test des rangs signés de Wilcoxon

Soient $(X_1, \dots, X_n)$ un n-échantillon d'une v.a. X absolument continue de médiane inconnue M et M_0 une valeur fixée. On suppose de plus que la distribution de X est symétrique (Médiane= Moyenne).

On veut tester

$$H_0: M = M_0 \quad vs \qquad H_1: M \neq M_0 \ \ (\ H_1: M > M_0 \quad ou \ \ H_1: M < M_0 \ \ ).$$

- On détermine les différences $D_i = X_i - M_0\,, i = \overline{1,n}$,
- On ordonne les $|D_i|$ de la plus petite à la plus grande,
- On calcule les rangs R_i^+ de $|D_i|, i = \overline{1,n}$,
- On associe à R_i^+ le signe de $D_i, i = \overline{1,n}$.

Sous, H_0, les différences D_i sont distribuées symétriquement autour de zéro :

$$F_D(-c) = P(D_i \leq -c) = P(D_i \geq c) = 1 - F_D(c).$$

On note W^+ la somme des rangs positifs (des rangs des écarts positifs) et W^- la somme des rangs négatifs (des rangs des écarts négatifs)

$$W^+ = \sum_{i=1}^{n} R_i^+ \, \mathrm{Z_i}, \qquad W^- = \sum_{i=1}^{n} R_i^+ \, (1 - \mathrm{Z_i})$$

avec

$$\mathrm{Z_i} = \mathbb{1}_{\{D_i > 0\}} = \begin{cases} 1 & \text{si } D_i > 0, \\ 0 & \text{si } D_i < 0. \end{cases}$$

Les variables aléatoires $\mathrm{Z_i}$, $i = \overline{1,n}$ sont indépendantes et les vecteurs $(R_1^+, \dots, R_n^+)$, $(\mathrm{Z_i}, \dots \mathrm{Z_n})$ sont indépendants

On a toujours $W^+ + W^- = \sum_{i=1}^{n} R_i^+ = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$. Il suffit donc de considérer W^+ (ou bien W^-).

- $0 \leq W^+ \leq \frac{n(n+1)}{2}$.

- Le cas $W^+ = 0$ correspond à tous les $D_i < 0$ $(X_i < M_0)$ et le cas $W^+ = \frac{n(n+1)}{2}$ correspond à tous les $D_i > 0$ $(X_i > M_0)$
- Inversement $W^- = 0$ correspond à tous les $D_i > 0$ et $W^- = \frac{n(n+1)}{2}$ correspond à tous les $D_i < 0$

Théorème :

Sous H_0, les statistiques W^+ et W^- ont la même distribution et sont des statistiques libres en loi. De plus, on a

$$E_{H_0}(W^+) = \frac{n(n+1)}{4}, Var_{H_0}(W^+) = \frac{n(n+1)(2n+1)}{24}$$

$$P_{H_0}(W^+ = k) = \frac{w_n(k)}{2^n}$$

où $w_n(k)$ est le nombre de manières d'affecter les signes plus et moins aux entiers $1, \dots, n$ tel que $W^+ = k$.

Preuve:

Sous H_0, la variable indicatrice $Z_i, i = 1, \dots, n$, suit une loi de Bernoulli de paramètre

$$p = P(Z_i = 1) = P(D_i > 0) = \frac{1}{2}.$$

Ainsi $(Z_i) = \frac{1}{2}$, $Var(Z_i) = \frac{1}{4}$ et

$$E(W^+) = \sum_{i=1}^{n} E(R_i^+ Z_i) = \sum_{i=1}^{n} E(Z_i)E(R_i^+) = \frac{1}{2}\sum_{i=1}^{n} E(R_i^+) = \frac{1}{2}\sum_{i=1}^{n} \frac{n+1}{2} = \frac{n(n+1)}{4}.$$

La distribution de W^+ dépend de celle de $(Z_1, \dots, Z_n)$ qui prend ses valeurs dans l'ensemble de tous les $n-$uplets $(z_1, \dots, z_n)$ avec composantes un ou zéro (dispositions ordonnées avec répétition de n éléments pris parmi un et zéro) qui sont au nombre de 2^n. Ces arrangements sont équiprobables sous H_0 , c.à.d.

$$P(Z_1 = z_1, \dots, Z_n = z_n) = \frac{1}{2^n}, \forall\ (z_1, \dots, z_n)\ arrangement\ possible.$$

La loi de W^+, sous H_0, est déterminée par énumération directe. Les valeurs de W^+sont calculées pour chaque disposition $(z_1, \dots, z_n)$ et

$$P(W^+ = k) = \frac{w_n(k)}{2^n}$$

où $w_n(k)$ est le nombre de manières d'affecter les signes plus et moins aux entiers $1, \dots, n$ (nombre de dispositions $(z_1, \dots, z_n)$) pour lesquelles que $W^+ = k$.

Exemple :

Dans le cas $n = 4$, on a $0 \leq W^+ \leq 10$ et $E(W^+) = 5$. Il ya $2^4 = 16$ manières d'affecter les signes plus et moins aux entiers 1,2,3,4 (16 dispositions possibles)

Disposition	Rangs	Valeur de W^+
(0,0,0,0)	0	0
(1,0,0,0)	1	1
(0,1,0,0)	2	2
(0,0,1,0)	3	3
(0,0,0,1)	4	4
(1,1,0,0)	1,2	3
(1,0,1,0)	1,3	4
(1,0,0,1)	1,4	5
(1,1,1,0)	1,2,3	6
(1,1,0,1)	1,2,4	7
(1,0,1,1)	1,3,4	8
(0,1,1,1)	2,3,4	9
(1,1,1,1)	1,2,3,4	10
(0,1,0,1)	2,4	6
(0,0,1,1)	3,4	7
(0,1,1,0)	2,3	5

Valeur de W^+ k	Fréquence $w_4(k)$	$P(W^+ = k)$
0	1	$\frac{1}{16} = 0.0625$
1	1	0.0625
2	1	0.0625
3	2	0.125
4	2	0.125
5	2	0.125
6	2	0.125
7	2	0.125
8	1	0.0625
9	1	0.0625
10	1	0.0625

Remarque :

La distribution de W^+, sous H_0, est symétrique par rapport à $E(W^+)$

$$P_{H_0}(W^+ - E(W^+) = w) = P_{H_0}(W^+ - E(W^+) = -w),$$

et $$P_{H_0}(W^+ \leq k) = P_{H_0}(W^+ \geq 2E(W^+) - k), \ \ 0 \leq k \leq E(W^+).$$

Région de rejet et p-valeur :

Sous H_0, $W^+ \ et \ W^-$ sont presque égales.

Si $H_1: M > M_0$, alors H_0 est rejetée pour les grandes (fortes) valeurs de W^+ (faibles valeurs de W^-).

Si $H_1: M < M_0$, alors H_0 est rejetée pour de faibles valeurs de W^+ (fortes valeurs de W^-).

Si $H_1: M \neq M_0$, alors H_0 est rejetée pour de faibles ou de fortes valeurs de W^+ .

Soit

$$W = \begin{cases} Min(W^+, W^-) \ si \ H_1: M \neq M_0, \\ W^- \qquad\qquad si \ H_1: M > M_0, \\ W^+ \qquad\qquad si \ H_1: M < M_0. \end{cases}$$

On rejette H_0 si $W \leq w_\alpha^*$ où w_α^* est la valeur critique pour un seuil α fixé et pour n fixée telle que

$$P_{H_0}(W \leq w_\alpha^*) \leq \alpha, w_\alpha^* \ se \ lit \ sur \ la \ table.$$

Notons que le niveau de signification d'un test bilatéral est égal à 2 fois le niveau de signification d'un test unilatéral. De plus la valeur critique w_α^* d'un test bilatéral de niveau de signification α est égale à la valeur critique $w_{\alpha/2}^*$ d'un test unilatéral à gauche de niveau de signification $\alpha/2$.

La $p-$valeur d'une valeur w prise par W^+ est :

$$\alpha_0 = \begin{cases} P_{H_0}(W^+ \geq w) & si \ \ H_1: M > M_0, \\ P_{H_0}(W^+ \leq w) & si \ H_1: M < M_0, \\ 2Min(P_{H_0}(W^+ \leq w), P_{H_0}(W^+ \geq w)) & si \ \ H_1: M \neq M_0. \end{cases}$$

On rejette H_0, si $\alpha_0 \leq \alpha$.

Exemple :

Le nombre d'animaux observés (durant une année) dans 12 postes est donné dans le tableau suivant

x_i	29	12	62	64	43	21	35	89	8	32	47	61

On veut tester au seuil 5%

$$H_0: M = 50 \; vs \; H_1: M \neq 50$$

x_i	29	12	62	64	43	21	35	89	8	32	47	61
$d_i = x_i - 50$	-21	-38	12	14	-7	-29	-15	39	-42	-18	-3	11
$z_i = \lvert d_i \rvert$	21	38	12	14	7	29	15	39	42	18	3	11
$z_{(i)}$	3	7	11	12	14	15	18	21	29	38	39	42
$r_i^+ = rg(\lvert d_i \rvert)$	8	10	4	5	2	9	6	11	12	7	1	3
$Signe(d_i)$	-	-	+	+	-	-	-	+	-	-	-	+

$$w^+ = 4 + 5 + 11 + 3 = 23$$

$$w^- = \frac{n(n+1)}{2} - 23 = \frac{12 * 13}{2} - 23 = 55$$

$$w^- = 8 + 10 + 2 + 9 + 6 + 12 + 7 + 1 = 55$$

$$w = Min(55,23) = 23$$

De la table on lit $w^*_{0.05} = 13 < 23 \implies on \; accepte \; H_0$

Pour un test unilatéral à droite $H_1: M > 50$

$$w^*_{0.05} = 17 < w^- = 55 \rightarrow on \; accepte \; H_0.$$

Pour un test unilatéral à gauche $H_1: M < 50$

$$w^*_{0.05} = 17 < w^+ = 23 \rightarrow on\ accepte\ H_0.$$

Pour la p −valeur, on a

$$\alpha_0 = \begin{cases} P_{H_0}(W^+ \geq 23) = 1 - P_{H_0}(W^+ \leq 22) = 1 - 0.102 = 0.898 & si\ H_1: M > 50, \\ P_{H_0}(W^+ \leq 23) = 0.117 & si\ H_1: M < 50, \\ 2Min\left(P_{H_0}(W^+ \leq 23), P_{H_0}(W^+ \geq 23)\right) = 0.234. & si\ H_1: M \neq 50. \end{cases}$$

Approximation normale :

Pour les grands échantillons ($n \geq 30$) et sous H_0, on peut utiliser la statistique

$$Z = \frac{W^+ - \mathrm{E}(W^+)}{\sqrt{\mathrm{Var}(W^+)}} = \frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \rightsquigarrow N(0,1).$$

La p-valeur α_0d'une valeur w observée de W^+ est telle que

$$\alpha_0 = \begin{cases} 2Min\left(P_{H_0}(W^+ \leq w), P_{H_0}(W^+ \geq w)\right) & si\ H_1: M \neq M_0, \\ P_{H_0}(W^+ \geq w) = 1 - \Phi\left(\frac{w - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right) & si\ H_1: M > M_0, \\ P_{H_0}(W^+ \leq w) = \Phi\left(\frac{w - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right) & si\ H_1: M < M_0. \end{cases}$$

Les régions critiques associées sont respectivement $\begin{cases} |Z| > z_{1-\alpha/2}, & si\ H_1: M \neq M_0 \\ Z > z_{1-\alpha}, & si\ H_1: M > M_0, \\ Z < z_{\alpha}. & si\ H_1: M < M_0. \end{cases}$

z_α : quantile d'ordre α de la loi $N(0,1)$.

- Avec la correction de continuité, la p-valeur pour une valeur observée w de W^+est donnée par

$$\alpha_0 = \begin{cases} P_{H_0}(W^+ \leq w) = \Phi\left(\dfrac{w + 0.5 - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}\right) \ si\ \ H_1 : M < M_0, \\ P_{H_0}(W^+ \geq w) = 1 - \Phi\left(\dfrac{w - 0.5 - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}\right) \ si\ \ H_1 : M > M_0, \\ 2Min\left(P_{H_0}(W^+ \leq w), P_{H_0}(W^+ \geq w)\right) \quad si\ \ H_1 : M \neq M_0. \end{cases}$$

Les régions de rejet correspondantes sont

$$\begin{cases} W^+ \leq z_\alpha \sqrt{\dfrac{n(n+1)(2n+1)}{24}} - 0.5 + \dfrac{n(n+1)}{4}, si\ H_1 : M < M_0, \\ W^+ \geq z_{1-\alpha} \sqrt{\dfrac{n(n+1)(2n+1)}{24}} + 0.5 + \dfrac{n(n+1)}{4}, si\ \ H_1 : M > M_0, \\ W^+ \leq z_{\alpha/2} \sqrt{\dfrac{n(n+1)(2n+1)}{24}} - 0.5 + \dfrac{n(n+1)}{4} \text{ ou } \ W^+ \geq z_{1-\alpha/2} \sqrt{\dfrac{n(n+1)(2n+1)}{24}} + 0.5 + \dfrac{n(n+1)}{4}, si\ \ H_1 : M \neq M_0. \end{cases}$$

Remarques :

1. Si certaines valeurs de l'échantillon sont égales à M_0 on les ignore et on considère celles qui sont différentes de M_0.
2. **Cas d'ex-aequos :** Si l'échantillon considéré présente des ex -aequos (valeurs égales), on attribue à celles-ci un rang moyen. Soit t_l le nombre de valeurs ex-aequos ayant le l^{ieme} rang. On calcule W^+ et W^-comme précédemment mais dans ce cas elles ne sont pas tabulées.

   Dans le cas des grands échantillons, on peut utiliser l'approximation normale avec une variance de W^+ corrigée. On remplace le terme $\frac{n(n+1)(2n+1)}{24} = \mathrm{Var}(W^+)$ par

$$\frac{n(n+1)(2n+1)}{24} - \sum_l \frac{t_l(t_l^2 - 1)}{48}.$$

Exemple

Soit $(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = (9,7,5,7,4,5,5)$. Le vecteur des rangs correspondant est $(7,5.5,3,5.5,1,3,3)$.

Il y a 4 rangs différents : 1, 3, 5.5 et 7 avec $t_1 = 1,\ t_2 = 3,\ t_3 = 2,\ t_4 = 1$.

Exemple

Le nombre d'animaux observés (durant une année) dans 35 postes est donné dans le tableau suivant

21	17	43	81	32	102	7	43	39	11	67	23	142
117	44	39	82	93	28	145	0	17	77	53	50	60
9	14	40	19	101	104	33	2	22.				

On veut tester au seuil 5%

$$H_0: M = 50\ vs\ H_1: M \neq 50$$

On obtient les différences suivantes

0	3	-6	-7	-7	-10	10	-11	-11	-17	17	-18	-22
-27	27	-28	-29	-31	31	32	-33	-33	-36	-39	-41	-43
43	-48	-50	51	52	54	67	92	95				

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	21	17	43	81	32	102	7	43	39	11	67	23	142	117
$d_i = x_i - 50$	-29	-33	-7	31	-18	52	-43	-7	-11	-39	17	-27	92	67
$\lvert d_i \rvert = z_i$	29	33	7	31	18	52	43	7	11	39	17	27	92	67
$z_{(i)}$	3	6	7	7	10	10	11	11	17	17	18	22	27	27
$r_i^+ = rg(z_i)$	16	20.5	3.5	17.5	11	30	25.5	3.5	7.5	23	9.5	13.5	33	32
$signe(d_i)$	-	-	-	+	-	+	-	-	-	-	+	-	+	+

15	16	17	18	19	20	21	22	23	24	25	26	27	28
44	39	82	93	28	145	0	17	77	53	50	60	9	14
-6	-11	32	43	-22	95	-50	-33	27	3	/	10	-41	-36
6	11	32	43	22	95	50	33	27	3	/	10	41	36
28	29	31	31	32	33	33	36	39	41	/	43	43	48
2	7.5	19	25.5	12	34	28	20.5	13.5	1	/	5.5	24	22
-	-	+	+	-	+	-	-	+	+	/	+	-	-

29	30	31	32	33	34	35
40	19	101	104	33	2	22
-10	-31	51	54	-17	-48	-28
10	31	51	54	17	48	28
50	51	52	54	67	92	95
5.5	17.5	29	31	9.5	27	15
-	-	+	+	-	-	-

La taille de l'échantillon est réduite à $n = 34$. On a $w^+ = 280.5,\ w^- = 314.5\ \ (w^+ + w^- = \frac{34\times35}{2} = 595)$et

$E_{H_0}(W^+) = \frac{34(35)}{4} = 297.5$. Le tableau ci-dessus contient 8 paires d'ex-æquos. Donc

$$Var(W^+) = \frac{n(n+1)(2n+1)}{24} - \sum_l \frac{t_l(t_l^2-1)}{48} = \frac{34(35)(69)}{24} - \frac{2(2^2-1)}{48}8 = 3420.25.$$

Comme $n = 34$est assez grand, on peut approximer la loi de W^+ par une loi normale

$$Z = \frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}} \rightsquigarrow N(0,1).$$

- Approximation normale sans correction de continuité

$$\alpha_0 = \begin{cases} 2Min\left(\mathrm{P}_{H_0}\left(W^+ \le w\right), \mathrm{P}_{H_0}\left(W^+ \ge w\right)\right) & \text{si} \quad H_1 : M \neq 50, \\ \mathrm{P}_{H_0}\left(W^+ \le w\right) = \Phi\left(\dfrac{w - E(W^+)}{\sqrt{\mathrm{var}(W^+)}}\right) & \text{si} \quad H_1 : M < 50, \\ \mathrm{P}_{H_0}\left(W^+ \ge w\right) = 1 - \Phi\left(\dfrac{w - E(W^+)}{\sqrt{\mathrm{var}(W^+)}}\right) & \text{si} \quad H_1 : M > 50, \end{cases}$$

$$\alpha_0 = \begin{cases} 2Min(\Phi(-0.29),\ 1 - \Phi(-0.29)) & \text{si} \quad \mathrm{H}_1 : M \neq 50, \\ \Phi(-0.29) & \text{si} \quad \mathrm{H}_1 : M < 50, \\ 1 - \Phi(-0.29) & \text{si} \quad \mathrm{H}_1 : M > 50, \end{cases}$$

$$\alpha_0 = \begin{cases} 2Min(1 - \Phi(0.29), \Phi(0.29)) & \text{si } \mathrm{H}_1 : M \neq 50, \\ 1 - \Phi(0.29) & \text{si } \mathrm{H}_1 : M < 50, \\ \Phi(0.29) & \text{si } \mathrm{H}_1 : M > 50, \end{cases}$$

$$\alpha_0 = \begin{cases} 2Min(0.3859\ ,\ 0.6141) = 0.7718 & \text{si } \mathrm{H}_1 : M \neq 50, \\ 0.3859 & \text{si } \mathrm{H}_1 : M < 50, \\ 0.6141 & \text{si } \mathrm{H}_1 : M > 50. \end{cases}$$

$\alpha_0 > \alpha$ donc on accepte H_0.

- Approximation normale avec correction de continuité

$$\alpha_0 = \begin{cases} 2Min\left(\mathrm{P}_{H_0}\left(W^+ \geq w\right); \mathrm{P}_{H_0}\left(W^+ \leq w\right)\right) & \text{si } \mathrm{H}_1 : M \neq 50 \\ \mathrm{P}_{H_0}\left(W^+ \geq w\right) = 1 - \Phi\left(\dfrac{w - 0.5 - E(W^+)}{\sqrt{\mathrm{var}(W^+)}}\right) & \text{si } \mathrm{H}_1 : M > 50 \\ \mathrm{P}_{H_0}\left(W^+ \leq w\right) = \Phi\left(\dfrac{w + 0.5 - E(W^+)}{\sqrt{\mathrm{var}(W^+)}}\right) & \text{si } \mathrm{H}_1 : M < 50 \end{cases}$$

$$\alpha_0 = \begin{cases} 2Min(1 - \Phi(-0.299); \Phi(-0.28)) & \text{si } \mathrm{H}_1 : M \neq 50 \\ 1 - \Phi(-0.299) & \text{si } \mathrm{H}_1 : M > 50 \\ \Phi(-0.28) & \text{si } \mathrm{H}_1 : M < 50 \end{cases}$$

$$\alpha_0 = \begin{cases} 2Min(\Phi(0.299); 1 - \Phi(0.28)) & \text{si } \mathrm{H}_1 : M \neq 50 \\ \Phi(0.299) & \text{si } \mathrm{H}_1 : M > 50 \\ 1 - \Phi(0.28) & \text{si } \mathrm{H}_1 : \mathrm{M} < 50 \end{cases}$$

$$\alpha_0 = \begin{cases} 0.7794 > 0.05 & \text{si } \mathrm{H}_1 : M \neq 50 \\ 0.6141 > 0.05 & \text{si } \mathrm{H}_1 : M > 50 \\ 0.3897 > 0.05 & \text{si } \mathrm{H}_1 : M < 50 \end{cases}$$

Donc on accepte H_0.

Test du signe et test des rangs signés pour données appariées

Définition:

Deux échantillons $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ sont dits appariés lorsque chaque observation X_i est associée à l'observation $Y_i, i = 1, \dots, n$ (appariés =associés par paires). C'est par exemple le cas lorsque les deux échantillons proviennent d'un même groupe de malades avant et après traitement. Deux échantillons appariés ont donc la même taille.

Soient un n-échantillon $\big((X_1, Y_1), \dots, (X_n, Y_n)\big)$ issu d'un couple de variables aléatoires continues (X, Y). On détermine les différences $D_i = X_i, -Y_i,\ i = 1, \dots, n$. On suppose que les D_i sont de fonction de répartition F_D $(D = X - Y)$ supposée continue. On effectue un test de signe ou un test des rangs singés sur la médiane M_D de la différence.

En générale, on teste $H_0: M_D = 0\ \ vs\ \ H_1: M_D \neq 0$ $\ (H_1: M_D > 0\ \ ou\ \ H_1: M_D < 0\ )$.

Exemple :

Soient les notes obtenues par 15 étudiants lors de deux tests

Etudiant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Test1(X)	8	7.5	9	15.5	12	13	6.5	5	12.5	11	3.5	14.5	10	8.5	7.5
Test2 (Y)	9	6.5	11.25	17.5	10.5	13.25	9.25	8.25	12.5	12.75	2.75	14	11.25	6	10

On pose $D = X - Y$ et on teste

$$H_0: M_D = 2\ \ vs\ \ H_1: M_D \neq 2.$$

d_i=$x_i - y_i$	-1	1	-2.25	-2	1.5	-0.25	-2.75	-3.25	0	-1.75	0.75	0.5	-1.25	2.5	-2.5
$signe\ (d_i - 2)$	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-

La statistique du test de signe est : $T^+ = \sum_{i=1}^{14} \mathbb{1}_{D_i>0}$ qui suit sous H_0 $une\ loi\ B(15,0.5)$.

On a $t^+ = 1$ et $\alpha_0 = P_{H_0}(T^+ \leq 1) = 0.0005$<0.05, donc on rejette H_0.