Chapter 1

Les tests non paramétriques

1.1 Introduction

Dans le chapitre précédent, nous avons vu qu'un test paramétrique nécessite que les données soient issues d'une distribution paramétrée; aussi les hypothèses nulle et alternatives du test portent sur un paramètre statistique (moyenne, variance ou proportion). Ces tests nécessitent des conditions de validité, notamment en ce qui concerne la taille de l'échantillon, la normalité de la distribution et l'égalité des variances.

Dans ce présent chapitre, nous allons traiter un autre type des tests qui sont moins exigeants par rapport à ces conditions, ces tests sont dits non paramétriques et ne font aucune hypothèse sur la distribution sous-jacente des données.

Lorsque les données sont quantitatives, les test non paramétriques transforment les données en rangs et mesurent l'accord entre les rangs observés et ce que devrait être ces rangs sous une hypothèse nulle.

Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.

Il existe une grande diversité de tests non paramétriques.

1. Variables qantitatives

Un échantillon	deux échantillons	plus de deux échantillons
-Tester les valeurs douteuses dans un échantillon: **Test de Dixon** -Tester si la valeur médiane de l'échantillon s'écarte d'une valeur théorique:**Test de Wilcoxon** -Tester si la distribution dans l'échantillons suit une loi donnée:**Test d'ajustement de Kolmogorov-Smirnov** cas particulier: si la distribution de l'échant suit une loi normale: **Test de normalité de Kolmogorov-Smirnov**	-Comparer la distribution de 2 échants non appariés: **1.Test de Kolmogorov Smirnov** **2.Test de Wilcoxon Mann-Whitney** -Comparer la distribution de échants appariés:**Test de Test de Wilcoxon apparié**	-Comparer la distribution de plusieurs échants Non appariés(indépts): **Test de Kruskal-Wallis** -Comparer la distribution de plusieurs échants appariés **Test de Friedman**

2. Variables qualitatives

Un échantillon	deux échantillons	plus de deux échantillons
-Tester si la distribution de l'échant suit une loi binomiale:**Test binomial**	-Comparer la distribution de 2 échantillons non appariés **Test de Khi2** -Comparer la distribution de 2 échantillons appariés: **Test de McNemar**	-Comparer la distribution de plusieurs échants non appariés:**Test de Khi2** -Comparer la distribution de plusieurs échants appariés: **Test de Cochran**

Les tests non paramétriques que nous allons aborder dans ce présent chapitre sont :

1. Comparaison de deux échantillons non appariés par un test de **Kolmogorov-Smirnov**.
2. Comparaison de deux échantillons appariés par un test de **Wilcoxon**.
3. Comparaison de plusieurs échantillons non appariés par un test de **Kruskal-Wallis**.
4. Comparaison de deux échantillons appariés (cas des variables qualitatives) par un test de **McNemar**.

1.2 Test de Kolmogorov-Smirnov

On considère deux échantillons indépendants X_1, X_2,..., X_n *iid* de fonction de répartition F_1 et Y_1, Y_2,..., Y_n *iid* de fonction de répartition F_2. On vise à savoir si les observations proviennent de la même population au regard de la variable d'intérêt; ou de manière équivalente : si la distribution de la variable d'intérêt est la même dans les deux échantillons.
Le test de Kolmogorov-Smirnov repose sur l'écart maximum entre les fonctions de répartition empiriques.

Définition 1 *: **La fonction de répartition empirique***
Soit $X = (X_1,..., X_n)$ un n-échantillon d'une loi de probabilité P absolument continue par rapport à la mesure de Lebesgue sue $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, et soit $x = (x_1,..., x_n)$ une observation de cet échantillon. La fonction de répartition associée à P, notée F et donnée par :
$\forall x \in \mathbb{R} : F(x) = P(X_i \leq x)$ *est inconnue.*
On peut estimer F par F_n; la fonction de répartition empirique associée à l'échantillon X définie comme suit :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq x} = \begin{cases} 0 \;\; si\, X < x_1 \\ \frac{1}{n} \;\; si\, x_i \leq X < x_{i+1} \\ 1 \;\; si\, X \geq x_n \end{cases}$$

.

Notons que pour chaque $x \in \mathbb{R}$ fixé, F_n est une variable aléatoire à valeurs dans $[0,1]$ et que par la loi forte des grands nombres, $F_n(x) \overset{p.s}{\to} F(x)$

1.2.1 Les hypothèses

Les hypothèses à tester sont :

H_0: Les deux échantillons se comportent de la même manière $F_1(x) = F_2(x)$

H_1: Les deux échantillons se comportent différemment $F_1(x) \neq F_2(x)$

1.2.2 Statistique du test de K-S

On note F_{n_1} la fonction de répartition associée au premier échantillon, F_{n_2} celle associée au deuxième échantillon. La statistique du test de K-S pour comparer deux échantillons indépendants est définie comme suit :

Cas des petits échantillons

- test bilatéral :

$$KS = D_{n_1,n_2} = \sup |F_{n_1}(x) - F_{n_2}(x)|$$

- test unilatéral

$$KS = D_{n_1,n_2} = \sup(F_{n_1}(x) - F_{n_2}(x))$$

Cas des grands échantillons

$KS = D_{n_1,n_2}\sqrt{\dfrac{n_1 n_2}{n_1+n_2}}$; $D_{n_1,n_2} = \sup |F_{n_1}(x) - F_{n_2}(x)|$

Dans ce cas la valeur seuil est lu sur la table suivante :

seuil α	0.1	0.05	0.025	0.01	0.005	0.001
KS_{crit}	1.22	1.36	1.48	1.63	1.73	1.95

1.2.3 Test de Wilcoxon apparié

C'est un test pour comparer deux échantillons aussi, mais cette fois ci les échantillons ne sont pas indépendants (ils sont appariés c-à-d dépendants).

L'appariement

Deux échantillons qui sont appariés c-à-d : qu'ils ne sont pas indépendants l'un de l'autre. Il peut s'agir par exemple d'une variable qui a été mesurée.

▷ Deux fois par le même opérateur dans des conditions de traitement différentes.

▷ Par deux opérateurs différents dans les même conditions de traitement.

▷ À deux instants t_1 et t_2 pour étudier l'effet d'un traitement au cours du temps.

Il peut s'agit aussi des observations obtenus chez les même individus.

▷ Étude avant/après (douleurs avant et après la prise d'un antalgique).

∘ *Exemple 1* : cas témoins (observations obtenues chez des individus différents présentant des caractères similaires).

∘ *Exemple 2* : cas témoins (appariement sur les facteurs de risque).

L'intérêt de l'appariement est qu'on peut contrôler la variabilité inter-individuelle et donc augmenter la puissance statistique dans la comparaison.

Principe du test de Wilcoxon

Nous disposons de 2 échantillons dépendants de même taille $n = n_1 = n_2$, et nous définissons le couple (X_1, X_2) pour chaque sujet.

Nous formulons une nouvelle variable : $D = X_1 - X_2$, variable aléatoire de moyenne μ.

Hypothèses à tester

Nous voulons tester les hypothèses suivants :

H_0 : "Les deux variables X_1 et X_2 ont la même distribution"

c-à-d fonctions de répartition identiques : $F_1(x) = F_2(x)$

H_1 : "X_1 et X_2 n'ont pas la même distribution de probabilité"

$$F_1(x) \neq F_2(x)$$

Nous procédons le test des rangs signé de Wilcoxon de la manière suivante :

1. On calcule les différences entre les valeurs de chaque couple de mesures : $d_i = x_{i_1} - x_{i_2}$.
2. On élimine les différences nulles.
3. On prend les valeurs absolues des différences $|d_i|$, mais on retient le signe de la différence.
4. On classe les $|d_i|$ de façon croissante.
5. On calcule R^+ la somme des rangs positifs : $R^+ = \sum_{i,d_i>0} r_i$ et R^- la somme des rangs négatifs : $R^- = \sum_{i,d_i<0} r_i$, sachant que la somme totale des rangs des écarts est égale à $\frac{n(n+1)}{2}$.
   On peut écrire : $R^- = \frac{n(n+1)}{2} - R^+$.
6. On prend : $w = \min(R^+, R^-)$

La statistique du test de Wilcoxon apparié noté w est donc définie comme suit :

$$w = \min(R^+, R^-)$$

La région critique du test

Fixant un seuil de signification α, on distingue les deux cas suivants :

- Si$n < 15$: petit échantillon

on rejette l'hypothèse nulle H_0 si :

$$R_+ \geq w_{crit}(+) \quad ou \quad R_- \leq w_{crit}(-)$$

où : $w_{crit}(\pm)$ sont les seuils critiques lues dans la table spécifique de Wilcoxon.

- Si $n \geq 15$: grand échantillon

Dans ce cas, on peut considérer une approximation de la loi normale et par conséquence on utilise la table des valeurs critiques z_α de la loi normale.

$$Z_{obs}(+) = \frac{(R^+ - m)}{\delta}$$
$$Z_{obs}(-) = \frac{(R^- - m)}{\delta}$$

avec : $m = \frac{n(n+1)}{4} \quad et \quad \delta^2 = \frac{n(n+1)(n+2)}{24}$

Et on compare $Z = \max(Z_{obs}(+), Z_{obs}(-))$ avec la valeur critique z_α.

- Si $Z \geq z_\alpha$: on rejette H_0 au niveau α.
- Si $Z < z_\alpha$: on accepte H_0 au niveau α.

Exemple 1

On dispose 13 clônes de peuplier dont on a mesuré la concentration en Aluminium dans le bois à deux instants différents (Août et Novembre) au sein d'une zone polluée.

clône	Aôut	Novembre
1	8.1	11.2
2	10	16.3
3	16.5	15.3
4	13.6	15.6
5	9.5	10.5
6	8.3	15.5
7	18.3	12.7
8	13.3	11.1
9	7.9	19.9
10	8.1	20.4
11	8.9	14.2
12	12.6	12.7
13	13.4	36.8

Peut-on observer entre Août et Novembre une différence de concentration en Aluminium dans le bois de peupliers ?

Solution

Le test de Wilcoxon apparié.

L'hypothèses du test :

$H_0 : F_1(x) = F_2(x)$ "il n'y a pas de différence entre Août et Novembre"

$H_1 : F_1(x) \neq F_2(x)$ "il y a une différence"

La statistique du test :

$$w = \min(R^+, R^-)$$

telle que

$$R^- = \sum_{i, d_i < 0} r_i$$

$$R^- = \sum_{i, d_i < 0} r_i$$

$d_i = N - A$	$\|d_i\|$	rang	rang(+)	rang(-)
3.1	3.1	6	6	-
6.3	6.3	9	9	-
-1.2	1.2	3	-	3
2	2	4	4	-
1	1	2	2	-
7.2	7.2	10	10	-
-5.6	5.6	8	-	8
-2.2	2.2	5	-	5
12	12	11	11	-
12.3	12.3	12	12	-
5.3	5.3	7	7	-
0.1	0.1	1	1	-
23.4	23.4	13	13	-
			$\sum$=75	$\sum$=16

On a : $R^+ = 75$, $R^- = 16$ et , $w_{crit}(-) = 17$

la règle de décision :

$R^- < w_{crit}$: on rejette H_0

On conclure que la concentration en Aluminium est différente entre les deux mois.

Exemple 2

On cherche à savoir si en entrainement régulier modifie la tension artérielle des personnes. On a recueilli la tension systolique de n=8 personnes dont on a appliqué un programme d'entraînement spécifique pendant 6 mois. Après l'entaînement, on a mesuré de nouveau la tension chez ces personnes.

Les résultats obtenus sont les suivants:

N	avant	après	d_i	$\|d_i\|$	rang	rang(+)	rang(-)
1	130	120	10	10	5	5	-
2	170	163	7	7	4	4	-
3	125	120	5	5	2	2	-
4	170	135	35	35	7	7	-
5	130	143	-13	13	6	-	6
6	130	136	-6	6	3	-	3
7	145	144	1	1	1	1	-
8	160	120	40	40	8	8	-
						$\sum$=27	$\sum$=9

Peut-on conclure qu'il y a une amélioration dans la tension artérielle chez ces personnes après les entraînement ?

L'hypothèses du test :

H_0 : "La tension est la même avant et après les entaînement" c-à-d $F_1(x) = F_2(x)$

H_1 : "La tension après est moins qu'avant" c-à-d $F_1(x) < F_2(x)$

La statistique du test :

On a $w = \min(R^+, R^-) = 9$ telle que : $R^+ = 27$, $R^- = 9$

la règle de décision :

On a $w_{crit} = 4$

On remarque que $R^+ > w_{crit}$ alors on rejette H_0 au seuil $\alpha = 5$

donc : la tension n'est pas la même avant et après l'entraînement, i.e : l'entraînement a un effet sur la tension.

1.2.4 Test de Kruskal-Walis

C'est un test pour comparer la distribution de $k > 2$ échantillons non appariés c-à-d vérifie si plusieurs échantillons appartiennent à la même population. Il s'agit de l'homologue non paramétrique de l'ANOVA à un facteur, mais avec le sérieux avantages de ne pas tenir

compte de la loi de la distribution de la variable étudiée ni l'égalité des variances entre les échantillons. Ce test est une extension généralisée du test de Wilcoxon-Mann-Whitney et par conséquence il fonctionne de la même façon en remplaçant les valeurs de la variable étudiée par leurs rangs.

On admet les notations suivants :

- k : le nombre totale d'échantillons.
- N : le nombre totale d'observations.
- n_i : le nombre d'observations dans l'échantillon i.
- r_i : la somme des rangs dans l'échantillon i.

La statistique du test

Soit $\bar{r}$ la moyenne globale des rangs et $\bar{r}_i$ la moyenne des rangs dans l'échantillon i.

La statistique de Kruskal-Walis est défini comme suit :

$$K = \frac{12}{N(N+1)} \sum_i n_i (\bar{r}_i - \bar{r})^2 ...(*)$$

Qui est bien l'expression d'une variabilité inter-classe c-à-d la dispersion des moyennes conditionnelles autours de la moyenne globale.

La formule $(*)$ est équivalente à la formule :

$$K = \frac{12}{N(N+1)} \sum_i \frac{r_i^2}{n_i} - 3(N+1)$$

La région critique

1) Pour des effectifs faible, on utilise la table de Kuskal-Walis pour déterminer la valeur critique k_{crit} à ne pas dépasser au seuil α.

$\rightarrow$ On rejette l'hypothèse nulle H_0 qui suppose que les échantillons proviennent des populations identiques si $K > k_{crit}$

2) Pour des effectifs suffisamment grands ($n_i > 5$), la statistique K suit approximativement une loi de khi-deux à (k-1) degré de liberté (χ^2_{k-1}) lorsque l'hypothèse H_0 est vrai.

$\rightarrow$ Dans ce cas, On rejette l'hypothèse nulle H_0 si $K > k_{crit} = \chi^2_{k-1_1-\alpha}$

Exemple

Soit 3 forages d'eau dont on a mesuré la concentration en Magnésium dans l'eau de manière quotidienne pendant 5 jours (mg/l).

forage 1	forage 2	forage 3
15	15	19
20	16	20
20	21	21
22	23	23
25	25	25

Peut-on observer une différence de concentration en Magnésium dans l'eau entre les trois forages?

Solution

Les hypothèses du test

H_0 : "Il n'existe pas une différence de concentration en 3 forages".

H_1 : "Il existe une différence de concentration en 3 forages".

N=15 , $n_i = 5$, k=3

	forage 1	r_1	forage 2	r_2	forage 3	r_3
	15	1.5	15	1.5	19	4
	20	6	16	3	20	6
	20	6	21	8.5	21	8.5
	22	10	23	11.5	23	11.5
	25	14	25	14	25	15
total	/	37.5	/	38.5	/	44
$\bar{r}_i$	/	7.5	/	7.7	/	8.8

On a : $\bar{r}$= 8

et

$$K = \frac{12}{N(N+1)} \sum_i n_i(\bar{r}_i - \bar{r})^2$$

$$K = 0.245$$

$K < k_{crit} = 5.78$, donc on accepte l'hypothèse nulle au seuil α=5.

par conséquent, il n'y a de différence de concentration entre les trois forages.