Tests statistiques (3): Tests non-paramétriques

A. Latouche

Tests Non-paramétrique

On considère 2 échantillons : $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$

- Ne nécessitent pas l'estimation de la moyenne (ou de la variance)
- Utilisent le **rang** des observations et pas la valeur des observations (x_i, y_j)
- Test de Wilcoxon pour 2 échantillons indépendants (Mann et Whitney) et Test de Wilcoxon (2 échantillons appariés)
- Pour $k > 2$ on utilisera le test de Kruskal et Wallis

Tests non-paramétrique

Lorsqu'on ne peut pas supposer que les variables sont normales et de même variance, on peut utiliser des tests dits non paramétriques qui sont valables quelles que soient les lois des variables de base.

Avantages des tests non paramétriques

1. Leur emploi se justifie lorsque les conditions d'applications des autres méthodes ne sont pas satisfaites, même après d'éventuelles **transformations** de variables.

2. Pour des échantillons de taille très faible jusqu'à $n = 6$, la seule possibilité est l'utilisation d'un test non paramétrique [1]

Remarques

1. Les tests paramétriques, quand leurs conditions sont remplies, sont plus **puissants** que les tests non-paramétriques.

[1] sauf si la distribution de la population est connue

Contexte

Objectif : Comparer les *distributions* de 2 échantillons dont les lois sont **inconnues**

- On observe les réalisations de 2 échantillon $(X_1, \ldots, X_{n_x})$ de loi P_x et $(Y_1, \ldots, Y_{n_y})$ de la loi P_y.

On souhaite tester l'hypothèse

$$H_0 : P_x = P_y$$

ce qui revient à tester

$$H_0 : F_1 = F_2$$

Rappel : La loi d'un variable aléatoire est complètement déterminée par sa fonction de répartition (ou sa densité)

Contexte

Si les observations correspondent

- à la pression artérielle (avec traitement) de n_x patients

- à la pression artérielle de n_y patients sans traitement

L'hypothèse

$$H_0 : P_x = P_y$$

correspond à une absence d'efficacité du traitement sur la pression artérielle

Test de Wilcoxon cas indépendants : Statistique de Mann-Whitney

2 échantillons indépendants $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_m)$
Ce test repose sur l'idée que si l'on mélange les 2 séries et qu'on ordonne le tout par valeurs croissantes on doit obtenir un mélange homogène

1. On ordonne les 2 suites, et on compte le nombre total de couples (x_i, y_i) où x_i à un rang plus grand que y_i
2. Soit U ce nombre [2]
   - U varie de 0 à $n \times m$
   - Si U=0 on est dans cette situation

$$x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$$

   - Si $U = n \times m$

$$y_1, y_2, \ldots, y_m, x_1, x_2, \ldots, x_n$$

[2] c'est la statistique de Mann-Whitney

Si les 2 distributions sont issues de la même loi

- $E(U) = \frac{nm}{2}$

- $Var(U) = \frac{nm(n+m+1)}{12}$

Pour n et m plus grands que 8 on peut utiliser une approximation normale

D'où

$$Z = \frac{U - nm/2}{\sqrt{Var(U}} \sim N(0,1)$$

La règle de décision est donc

On rejète $H_0 : F_1 = F_2$ si $|Z| > z_{\alpha/2}$ pour un α donné

Test de Wilcoxon cas indépendants : Statistique de Wilcoxon

Une autre approche (plus rapide) consiste à

- calculer la somme des rangs des individus de l'un des 2 groupes (le premier par exemple X)

- Soit W_X cette somme.

Sous $H_0 : F_1 = F_2$, on montre que

$$W_X = nm + \frac{n(n+1)}{2} - U$$

et $E(W_X)) = \frac{n(n+m+1)}{2}$ et $var(W_X) = \frac{nm(n+m+1)}{12}$

Illustration

On dispose de 2 groupes de souris d'effectifs 6 et 5. On mesure le même facteur quantitatif pour chaque souris.

Groupe 1	11	21	25	52	71	79
Groupe 2	22	43	72	91	116	

Les 2 groupes sont-ils différents significativement ?

Illustration

On classe toutes les observations par ordre croissant

11 21 22 25 43 52 71 72 79 91 116

On calcule sur cette répartition pour chaque valeur du groupe 1, le nombre de valeurs du groupe 2 qui lui est inférieur.

11 21 22 25 43 52 71 72 79 91 116

Illustration

On classe toutes les observations par ordre croissant

11	21	22	25	43	52	71	72	79	91	116

On calcule sur cette répartition pour chaque valeur du groupe 1, le nombre de valeurs du groupe 2 qui lui est inférieur.

11	21	22	25	43	52	71	72	79	91	116
0										

Illustration

On classe toutes les observations par ordre croissant

11	21	22	25	43	52	71	72	79	91	116

On calcule sur cette répartition pour chaque valeur du groupe 1, le nombre de valeurs du groupe 2 qui lui est inférieur.

11	21	22	25	43	52	71	72	79	91	116
0	0									

Illustration

On classe toutes les observations par ordre croissant

11	21	22	25	43	52	71	72	79	91	116

On calcule sur cette répartition pour chaque valeur du groupe 1, le nombre de valeurs du groupe 2 qui lui est inférieur.

11	21	22	25	43	52	71	72	79	91	116
0	0		1							

Illustration

On classe toutes les observations par ordre croissant

11	21	22	25	43	52	71	72	79	91	116

On calcule sur cette répartition pour chaque valeur du groupe 1, le nombre de valeurs du groupe 2 qui lui est inférieur.

11	21	22	25	43	52	71	72	79	91	116
0	0		1		2					

Illustration

On classe toutes les observations par ordre croissant

11	21	22	25	43	52	71	72	79	91	116

On calcule sur cette répartition pour chaque valeur du groupe 1, le nombre de valeurs du groupe 2 qui lui est inférieur.

11	21	22	25	43	52	71	72	79	91	116
0	0		1		2	2				

Illustration

On classe toutes les observations par ordre croissant

11	21	22	25	43	52	71	72	79	91	116

On calcule sur cette répartition pour chaque valeur du groupe 1, le nombre de valeurs du groupe 2 qui lui est inférieur.

11	21	22	25	43	52	71	72	79	91	116
0	0		1		2	2		3		

On somme tous ces nombres; ici : $U = 8$

Illustration

```
grp1=c(11,21,25,52,71,79)
grp2=c(22,43,72,91,116)
wilcox.test(grp1,grp2)
Wilcoxon rank sum test
data: grp1 and grp2
W = 8, p-value = 0.2468
```

$\Rightarrow$ On ne rejette pas H_0

Illustration

On veut comparer les performances de 2 groupes d'élèves (en minutes) à un test.

Groupe 1	22	31	14	19	24	28	27	28		
Groupe 2	25	13	20	11	23	16	21	18	17	26

Illustration

On veut comparer les performances de 2 groupes d'élèves (en minutes) à un test.

Groupe 1	22	31	14	19	24	28	27	28		
Groupe 2	25	13	20	11	23	16	21	18	17	26

- On réordonne les observations par ordre croissant (Groupe 1 souligné)

Illustration

On veut comparer les performances de 2 groupes d'élèves (en minutes) à un test.

Groupe 1	22	31	14	19	24	28	27	28		
Groupe 2	25	13	20	11	23	16	21	18	17	26

- On réordonne les observations par ordre croissant (Groupe 1 souligné)

11	13	$\underline{14}$	16	17	18	$\underline{19}$	20	21	$\underline{22}$	23	$\underline{24}$	25	26	$\underline{27}$	$\underline{28}$	$\underline{28}$	$\underline{31}$
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Illustration

On veut comparer les performances de 2 groupes d'élèves (en minutes) à un test.

Groupe 1	22	31	14	19	24	28	27	28		
Groupe 2	25	13	20	11	23	16	21	18	17	26

- On réordonne les observations par ordre croissant (Groupe 1 souligné)

11	13	$\underline{14}$	16	17	18	$\underline{19}$	20	21	$\underline{22}$	23	$\underline{24}$	25	26	$\underline{27}$	$\underline{28}$	$\underline{28}$	$\underline{31}$
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

$$W_X = 3 + 7 + 10 + 12 + 15 + 16 + 17 + 18 = 98$$

Sous H_0, on a

- $E(W_X) =$

Sous H_0, on a

- $E(W_X) = \dfrac{8(8+10+1)}{2} = 76$
- $Var(W_X) =$

Sous H_0, on a

- $E(W_X) = \dfrac{8(8+10+1)}{2} = 76$
- $Var(W_X) = \dfrac{8 \times 10(8+10+1))}{12} = 126.7$

Sous H_0, on a

- $E(W_X) = \dfrac{8(8+10+1)}{2} = 76$
- $Var(W_X) = \dfrac{8 \times 10(8+10+1))}{12} = 126.7$

La statistique de test observée vaut $\dfrac{98-76}{11.25} = 1.96$

- au risque $\alpha = 0.05$, on ne rejette pas H_0

- au risque $\alpha = 0.1$, on rejette H_0

Sous H_0, on a

- $E(W_X) = \dfrac{8(8+10+1)}{2} = 76$
- $Var(W_X) = \dfrac{8 \times 10(8+10+1))}{12} = 126.7$

La statistique de test observée vaut $\dfrac{98-76}{11.25} = 1.96$

- au risque $\alpha = 0.05$, on ne rejette pas H_0

- au risque $\alpha = 0.1$, on rejette H_0

$\bar{x}_1 = 24.13$ et $\bar{x}_2 = 19$

Sous H_0, on a

- $E(W_X) = \dfrac{8(8+10+1)}{2} = 76$
- $Var(W_X) = \dfrac{8 \times 10(8+10+1))}{12} = 126.7$

La statistique de test observée vaut $\dfrac{98-76}{11.25} = 1.96$

- au risque $\alpha = 0.05$, on ne rejette pas H_0

- au risque $\alpha = 0.1$, on rejette H_0

$\bar{x}_1 = 24.13$ et $\bar{x}_2 = 19$
On peut donc conclure à une plus grande rapidité du groupe 2 (au risque $\alpha = 0.1$)

Test de Wilcoxon pour données Appariées

On dispose de 2 mesures sur une unité statistique (pas d'indépendance entre les mesures)
Principe

1. On classe par ordre de valeurs absolue croissante les différences

2. On calcule ensuite la somme des rangs des différences positives (W_+)

On montre que

- $E(W_+) = \frac{n(n+1)}{4}$

- $Var(W_+) = \frac{n(n+1)(2n+1)}{24}$

A partir de n=10 on utilisera l'approximation normal

Test de Wilcoxon pour données appariées : Illustration

L'échelle de Hamilton est employée pour évaluer la dépression de 9 patients Avant et Après traitement : on est donc dans un cas **apparié**)

Avant	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
Après	0.88	0.65	0.60	2.05	1.06	1.29	1.06	3.14	1.29

`wilcox.test(x, y, paired = TRUE)`
V=40
ou
`wilcox.test(y - x)`
V=5
Les 2 tests fournissent la même p-value=0.03906

On rejette l'hypothèse d'égalité des distributions

Test de Wilcoxon pour données appariées : Illustration

Diff	0.952	-0.147	1.022	0.430	0.620	0.590	0.490	-0.080	0.010
Rang	8	3	9	4	7	6	5	2	1
Rang signé	8	-3	9	4	7	6	5	-2	1

Somme des rangs positifs= 40
Somme des rangs négatifs=5

Illustration: Pollution

On veut évaluer l'impact de l'interdiction de circulations dans certaines rues

Lieu	Sans	Avec
1	214	159
2	159	135
3	169	141
4	202	101
5	103	102
6	119	168
7	200	62
8	109	167
9	132	174
10	142	159
11	194	66
12	104	118
13	219	181
14	119	171
15	234	112

Calculer W_+, $E(W_+)$ et $Var(W_+)$

Illustration: Pollution

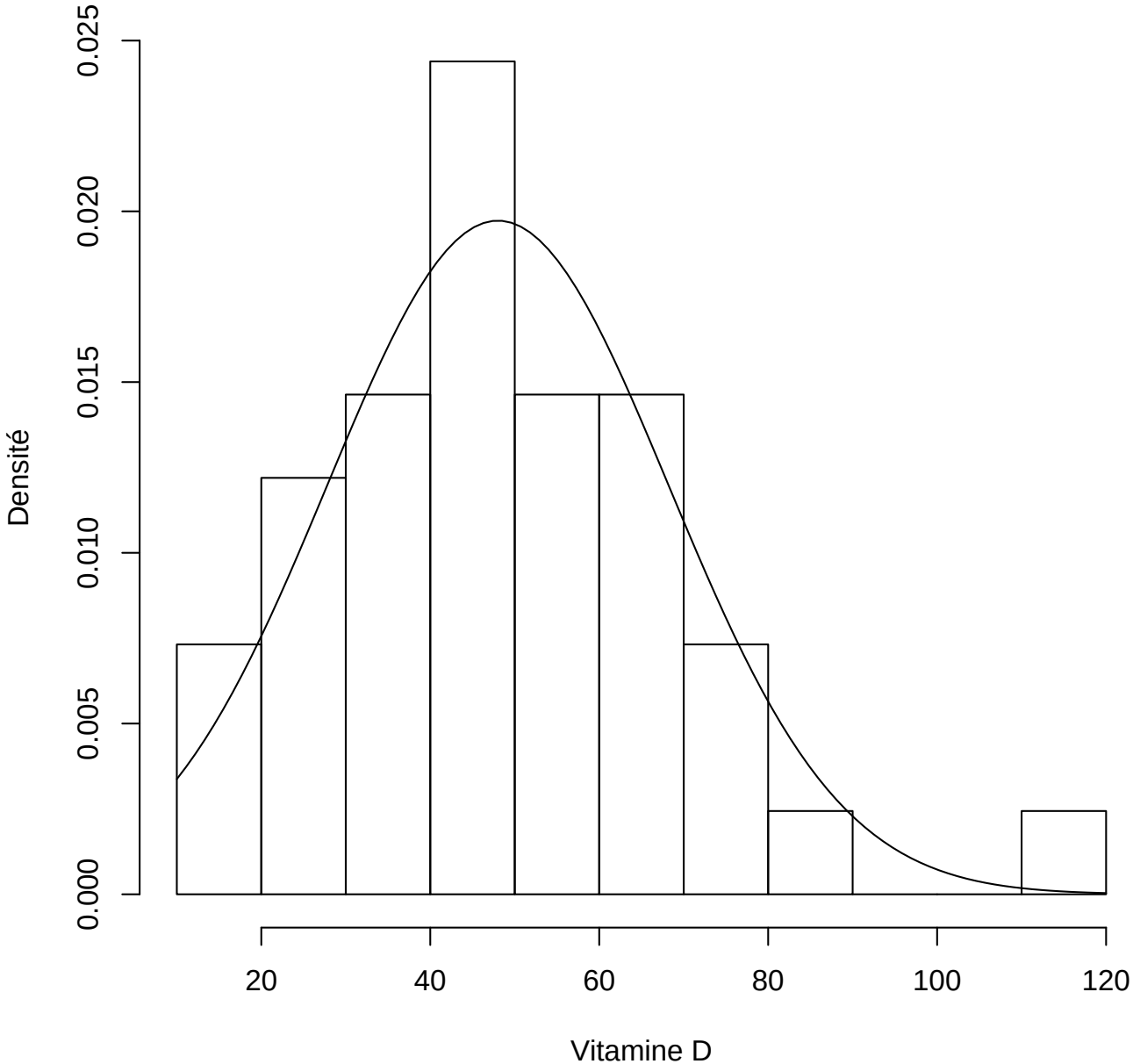
Lieu	Sans	Avec	diff
1	214	159	55
2	159	135	24
3	169	141	28
4	202	101	101
5	103	102	1
6	119	168	-49
7	200	62	138
8	109	167	-58
9	132	174	-42
10	142	159	-17
11	194	66	128
12	104	118	-14
13	219	181	38
14	119	171	-52
15	234	112	122

Test de la normalité

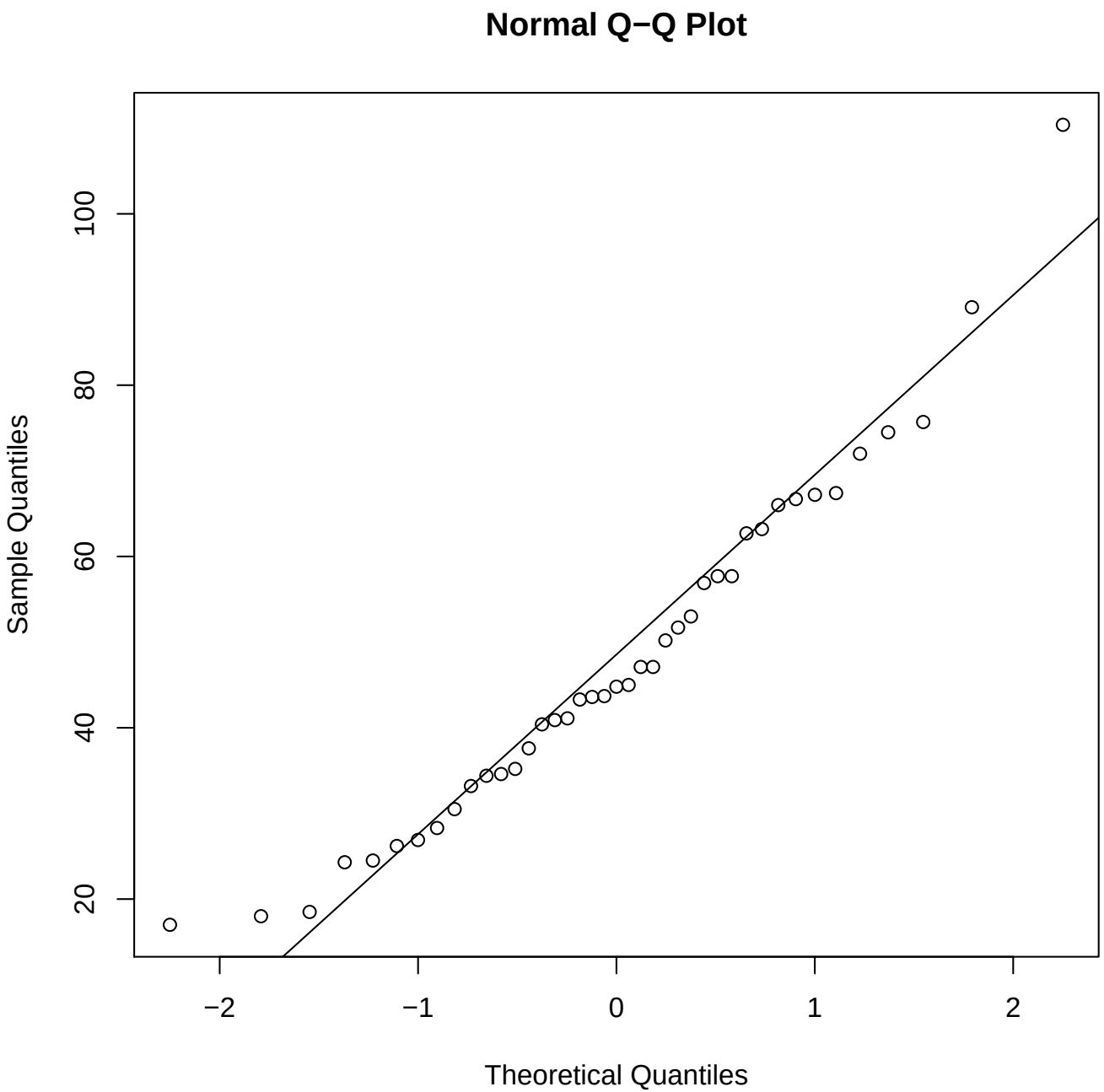
- Histogramme, Q-Q plot, droite de Henry

- Test de Shapiro-Wilks

Cours 4 : Densité de la loi $N(48, 20.22^2)$



Quantile-Quantile plot : Vitamine D



Pour construire un Q-Q plot sur un échantillon de données

(1) Classer les observations par ordre croissant

(2) Déterminer le percentile de chaque observation

(3) Identifier la valeur correspondant à chaque percentile (table de la loi normale, Z score)
Par exemple le quantile d'ordre 0.99 % vaut -2.33

(4) Faire un graphique des quantiles théoriques de la loi normale et des quantiles observés

Si les observations proviennent d'une loi normale on obtient une droite (de Henry)

Le carré d'un loi Normale