République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

ATMANI Fatima

Titre :

Sur quelques tests non paramétriques et applications

Membres du Comité d'Examen :

Pr. BENATIA Fatah	UMKBiskra	Président
Pr. YAHIA Djabrane	UMKBiskra	Encadreur
Dr. SOLTANE Louiza	UMKBiskra	Examinatrice

Septembre 2020

Dédicace

Je dédie ce humble travail à

mes chers parents.

REMERCIEMENTS

J'aimerais en premier lieu remercier mon dieu ***Allah*** *qui m'a donné la volonté et le courage pour la réalisation de ce travail*

J'exprime mes profondes gratitudes à mes parents

Je tiens à exprimer ma profonde gratitude à mon encadreur le Professeur **YAHIA Djabrane** *et je le remercie aussi de ses remarques importantes et de son suivi permanent de mon travail*

Je veux exprime aussi tout mon respect aux membres du jury, qui ont acceptés d'évaluer et de juger mon travail

Mes remerciements vont aussi à tous les enseignants du département de Mathématiques qui ont contribué à ma formation

toutes mes amies et toute personne qui ma aidée à la réalisation de ce travail

Fatima.

Table des matières

Table des figures

Liste des tableaux

Introduction

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses à la vue des résultats d'un échantillon, en quantifient le risque associé à la décision prise.

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, qui vient confirmer même lorsque les conditions d'applications des tests paramétriques sont réunies. L'avantage de ces derniers par rapport aux tests non paramétriques n'est pas transcendant.

Lorsque les effectifs sont faibles, les tests paramétriques à moins vraiment que l'hypothèse de normalité ne soit établie, ne sont pas bons, à la différence des tests non paramétriques, ces derniers sont alors sans concurrence. De plus, lorsque la variable d'intérêt est gaussienne, l'efficacité relative asymptotique du test de Wilcoxon-Mann-Withney (test non paramétrique) par rapport au test de Student (test paramétrique) est 95%, c-à-d lorsque l'hypothèse alternative est vraie, les moyennes sont différentes, s'il faut 95 observations pour aboutir à cette conclusion avec le test de Student, il en faudra 100 avec le test non paramétrique.

Des tables statistiques spécifiques pour les tests non paramétriques sont disponibles pour réaliser des tests exacts. Additionnée à cela, la convergence vers les lois asymptotiques est trés rapide. Dans la pratique, dès que les effectifs atteignent au niveau modéré (de l'ordre de 20 à 30 observations), les approximations sont déjà pleinement efficace. On prend le test de Wilcoxon–Mann-Whitney par exemple, il suffit que $n_1 > 10$ (ou $n_2 > 10$) pour que l'approximation à une loi normale soit valable.

Ce mémoire est un aperçu sur les tests non paramétriques, en focalisant sur les tests d'homogénéité entre deux distributions ou plus. Le travail est composé en deux chapitres comme suit :

Dans le premier chapitre nous rappelons un certain nombre de généralités autour des tests d'hypothèses et énonçons aussi des propriétés fondamentales sur quelques tests comme le test de Kolmogorov-Smirnov sur un seul échantillon, le test d'ajustement du khi-deux, le test de χ^2 d'indépendance et le test de Lilliefors.

Nous intéressons dans le deuxième chapitre sur les tests d'homogénéité entre deux distributions ou plus, comme le test de Wilcoxon-Mann-Whitney, test de Kruskal-Wallis, test des blocs de Wald-Wolfowitz, test de χ^2 d'homogénéité, test de Kolmogorov-Smirnov d'homogéniété et le test de Cramer-Von-Mises.

A l'aide du logiciel d'analyse statistique R, plusieurs exemples d'applications seront traités dans ce mémoire.

Chapitre 1

Généralités sur les tests non paramétriques

1.1 Notions générales

Consédirons une population comme étant l'ensemble Ω fini d'éléments ω_i appelés les individus, et on définit $X = (X_1, ..., X_n)$ l'ensemble d'individus (échantillon) issus de cette population.

1.1.1 Test statistique

Un test statistique (ou test d'hypothèse) est une **procédure de décision** qui permet de choisir entre deux hypothèses sur une population statistique après l'observation d'un échantillon. Un test statistique est cernée présenter par exemple : la répartition de l'age des poissions, la moyenne, la médiane, une fréquence,l'indépendance entre deux variables, ou bien l'homogènieté... Un test statistique peut être utilisé aussi pour vérifier le succés ou l'échec d'un événement.

1.1.2 Hypothèses nulle et alternative

Les hypothses nulle et alternative sont deux déclarations s'excluant mutuellement sur une population. Un test d'hypothèse utilise des données d'échantillon pour rejetté ou non l'hypothèse nulle. Dans le cas d'un test paramétrique on distingne deux types d'hypothèses : les hypothèses simples et les

hypothèses composés (multiples).

- Une hypothèse est dite simple si elle a la forme "$\theta = \theta_0$"ou $\theta \in \Theta$, shachant que Θ est un ensemble de toutes les valeurs de θ.

- Une hypothèse est dite composée si elle a la forme "$\theta \in A$" où $A \in \Theta$ ayant deux éléments ou plus.

1.1.3 Erreurs et risques

Lors de la prise de décision (choisir H_0 ou H_1) quatre situations sont possibles :

- Accepter H_0 et elle est vraie (bonne décision)

- Rejeter H_0 et elle est fausse (bonne décision)

- Rejeter H_0 et elle est vraie (appellé erreur de première espèce)

- Accepter H_0 et elle est fausse (appellé erreur de deuxième espèce)

Le risque est la probabilité de l'erreur, on a donc deux types de risque : risque de premier espèce et risque de deuxième espèce, le premier est généralement noté par "α" et le second "β".

$$\alpha = \alpha(\theta) = P(rejeter\ H_0 | H_0\ est\ vraie) = P(H_1/H_0).$$

$$\beta = \beta(\theta) = P(accepter\ H_0 | H_0\ est\ fausse) = P(H_0/H_1)$$

Les tests sont fondés sur le schéma suivant :

décision | vraie	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

1.1.4 Région de rejet - d'acceptation

La région du rejet d'un test est l'ensemble des observations $(x_1, ..., x_n)$ dans $\mathbb{R}^n$ pour le quel l'hypothèse nulle H_0 est écartée au profite de l'alternative H_1, on note généralement par W, et elle est déterminée

par la relation : $P\ (W/H_0) = \alpha$.

Le complémentaire de W est la région d'acceptation, on la note par $\bar{W}$, elle est définie par :

$$P(\bar{W}/H_1) = 1 - \alpha. \tag{1.1}$$

1.2 Tests paramétriques et non paramétriques

Lorsque l'on réalise des comparaisons de population ou que l'on copmare une population à une valeur théorique, il existe deux grands familles des tests : tests paramétriques,tests non paramétriques, le but des tests paramétriques de montrer une égalité sur certaines paramétriques, et bien sûr sur les conditions d'application du test vérifiées et par exemple d'un test paramétrique : la moyenne (test de comparaison de deux moyennes ou plus), ou test de comparaison de deux variances ou plus,test de Student, ect.

Par ailleurs les tests non paramétriques permettent **sans aucune hypothèse** sur la loi de probabilité (**distribution free**) suivie par la variable aléatoire impliquée de livrer des conclusions intéressantes sur l'étude de séries indépendantes comme sur celle de séries appariées, nous citerons simplement certains d'entre eux. On va aborder maintenant quelques types des tests non paramétriques :

1.2.1 Test de Kolmogorov-Smirnov (K-S)

L'idée du test est de *comparer la fonction de disribution empirique à la foncction de répartition*. Le test de K-S permet de tester n'importe quelle distribution.

Soit $X = (X_1, X_2, ...X_n)$ un n-échantillon d'une loi P **absolument continue** par rapport à la mesure de Lebesgue sur $(\mathbb{R}, \beta(\mathbb{R}))$ inconnue, soit $x = (x_1, ..., x_n)$ une observation de cet échantillon, et on

note F_n la distribution empirique associée à l'échantillon X définie pour $t \in \mathbb{R}$ par :

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i \leq t)} = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_{(i)} \leq t)}$$
$$= \begin{cases} 0, & x_{(i)} > t \\ \frac{i}{n}, & x_{(i)} \leq t < x_{(n)} \\ 1, & x_{(n)} \leq t \end{cases}$$

où les $x_{(i)}$ sont les statistiques d'ordre de l'échantillon (valeurs de l'échantillon rangées par ordre croissant). En d'autres termes, $F_n(t)$ est la proportion d'éléments de l'échantillon qui sont inférieurs ou égaux à x.

Propriété 1.2.1 *On a,*

· pour chaque $t \in \mathbb{R}$ fixé, $F_n(t)$ est une variable à valeur dans $[0, 1]$.

· si on recherche à tester l'hypothèse : $H_0(P = P_0) \Leftrightarrow H_0(F = F_0)$ où F_0 est la fonction de répartition de la loi normale.

·*Le théorème de Glivento – Gantelli* donne :

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \underset{n \to +\infty}{\longrightarrow} 0 \quad Ps. \tag{1.2}$$

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

$$KS = D_{KS}(P, P_0) = \sup_{t \in \mathbb{R}} |Fn(t) - F(t)|. \tag{1.3}$$

Proposition 1.2.1 *Si $(x_{(1)}, ..., x_{(n)})$ est la statistique d'ordre associée à l'échantillon X alors :*

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

• La région critique : on rejette H_0 si $D_{KS} > d_{n,\alpha}$, où $d_{n,\alpha}$ est le quantile théorique lu à partir la table de Kolmogorov.

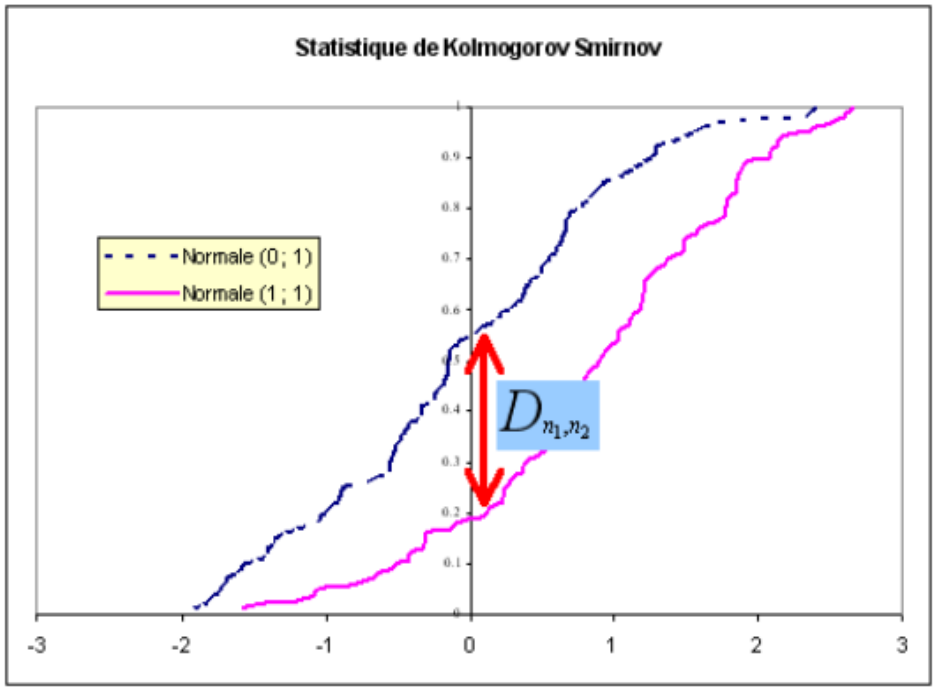

FIG. 1.1 – Détermination de la statistique de Kolmogorov-Smirnov

Exemple 1.2.1 *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 :

$x2 < -c(3.5, 16, 18, 14, 26, 17.5, 12, 22.5, 36, 10)$

$ks.test(x2, "ppois", lambda = 1/5)$

$One - sample\ Kolmogorov - Smirnov\ test$

$data : x_2$

$D = 0.88248, p - value = 3.442e - 07$

$alternative\ hypothes\ is : two - sided$

$p - value = 0.003$

Comme la p-valeur est inférieure à la valeur de α, 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%.

1.2.2 Tests de Pearson (Khi-2)

1.2.3 Distance de Khi-2

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 :

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

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 δ_* est la masse de Dirac.

Définition 1.2.1 *La distane de Khi-2 (Pearson, 1900) entre P et P_n est*

$$Q = D(P, P_n) = \sum_{k=1}^{m} \frac{(N_n - nP_k)^2}{nP_k},$$

avec $Q \sim \chi^2_{(m-1)}$ quand $n \to \infty$.

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 sur θ 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 χ^2 à $(m-1)$ degrés de liberté.

• La statistique de khi-2 définie par :

$$Q = D(P_n, P_0) = \frac{\sum (N_k - np_{0k})^2}{np_{0k}}. \tag{1.4}$$

• La région critique : on rejette H_0 si

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

Remarque 1.2.1 *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 $Q > \chi^2_{(m-r-1)}(1-\alpha) = q_{\alpha r}$.*

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_1, X_2 deux variables qualitatives telle que X_1 est à valeur dans $\{a_1, ..., a_m\}$ et X_2 à valeur dans $\{b_1, ..., b_n\}$. Sous H_0, la distribution de X_1 devrait *être indépendante* de celle de X_2. Par contre, si la disrtibution de X_1 est *liée* à celle de X_2, on rejette H_0 au profit de H_1, les deux variables X_1 et X_2 sont liées. Ainsi, on cherche à tester :

$$\begin{cases} H_0 : \text{Les variables } X_1, X_2 \text{ sont indépendantes} \\ H_1 : \text{Les variables sont liées} \end{cases}$$

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

$$Q_{ind} = \sum_{i=1}^{m}\sum_{j=1}^{l} \frac{(\frac{N_{i*}*N_{*j}}{n} - N_{ij})^2}{\frac{N_{i*}*N_{*j}}{n}}, \tag{1.5}$$

N_{i*} : Nombre de X_1 de valeurs a_i $(i = 1, ..., m)$.

N_{*j} : Nombre de X_2 de valeurs b_j $(j = 1, ..., k)$.

N_{ij} : Nombre de (X_1, X_2) de valeurs (a_i, b_j).

• La région critique :pour un risque de première espèce α, 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 χ^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_1, X_2.

Exemple 1.2.2 *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

TAB. 1.1 – Liaison entre les notes des étudiants et leurs sexe

Sous R, on utilisons les commandes associées sont données par $(k = 2, h = 3)$

$> A = matrix(c(25, 14, 15, 8, 5, 7), nrow = 2, byrow = T)$

$> chisq.test(A)\$p.value$

[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 (F,G).

Remarque 1.2.2 *On peut aussi considérés des caractères quantitatifs avec des valeurs réparties dans quelques intervalles disjoints appelés classes. Dans ce cas, on les traite caractères qualitatifs et leurs classes joueront le rôle de modalités.*

Test de Lilliefors

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

• La statistique du test est :

$$\begin{aligned} L_n &= \sqrt{n} KS \\ &= \max_{1\leq i\leq n} \max \left\{ \left| F_0(\frac{x_{(i)} - \overline{x}}{s_x}) - \frac{i}{n} \right|, \left| F_0(\frac{x_{(i)} - \overline{x}}{s_x}) - \frac{i-1}{n} \right| \right\} \end{aligned}$$

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

• La région critique : 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 conclure l'acceptation de H_0 comme suit :

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

$lillie fors(kolmogorov - smirnov) normality.test$

$data : rnorm(100, mean = 5, sd = 3)$

$D = 0.0646$, $p - value = 0.3841$.

Chapitre 2

Tests d'homogénéité

Après avoir rappelé quelques notions et concepts fondamentales de certains tests non paramétriques qui compare la fonction de distribution empirique à la fonction de répartition, on va aborder dans ce chapitre au test d'homogénéité entre deux distributions ou plus.

2.1 Test de Wilcoxon-Mann-Whitney

Certainement ce test est le plus populaire des tests non paramétriques. Il recouvre en réalité deux formulations sont équivalentes (ils peuvent se déduire l'un de l'autre), d'une part le test de Wilcoxon, d'autre part le test de Mann-Whitney.

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 les distributions de X et de$Y$ (d'autre façon étudier l'homogénéité). On dispose d'échantillons indépendantes : cas le plus habituel, ils ont été obtenus par tirage au sort dans deux populations différentes :

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

On considère deux échantillons indépendants $(X_1, ..., X_n)$ de la fonction de répartition F_1 et $(Y_1, ..., Y_m)$

aussi de même fonction de répartition F_2, on veut tester maintenent

$$\begin{cases} H_0 : F_1(X) = F_2(X+\theta), \theta = 0 \quad distributions\ identiques \\ H_1 : F_1(X) = F_2(X+\theta), \theta \neq 0 \quad distributions\ diff\acute{e}rents \end{cases}$$

θ paramètre de translation : décalage entre les fonctions de répartitions, on suppose que F_1 et F_2 sont continues.

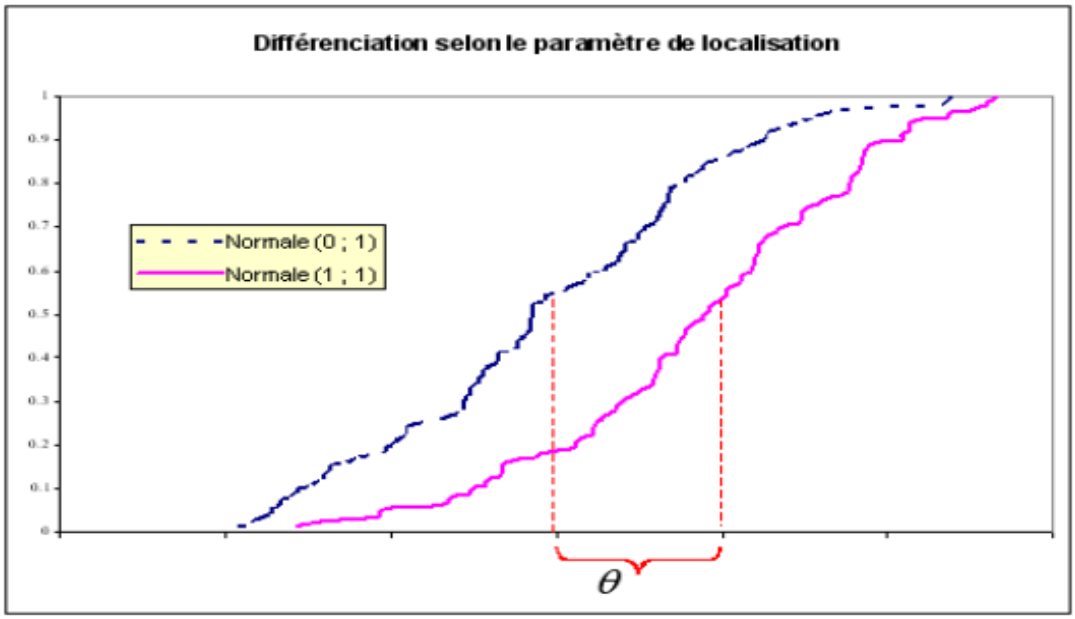

FIG. 2.1 – Paramètre de translation - Décalage entre 2 fonctions de répartition

La procudure du Wilcoxon-Mann-Whitney 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) où, $W = \sum_{i=1}^{N} R(X_i)$ telle que $R(X_i)$ le rang de la $i^{ème}$ observation de X.
4. calculer la statistique $U_{n,m}$.

La statistique de **Mann** et **Whitney** utilise la somme des rangs. Nous retrouvons bien l'idée de

décalage entre les distributions basé sur leur localisation, pour le test :

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases} \iff \begin{cases} H_0 : F_1(X) = F_2(X+\theta) \\ H_1 : F_1(X) = F_2(X+\theta) \end{cases} \tag{2.1}$$

Nous calculons les quantiles :

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = W_2 - \frac{n_2(n_2+1)}{2}.$$

Le principe de **Mann-Whitney** correspond à la plus petite quantité,

$$U = \min(U_1, U_2)$$

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

$$E(U) = \frac{n_1 n_2}{2}, \quad V(U) = \frac{(n_1+n_2+1)n_1 n_2}{12}. \tag{2.2}$$

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.

On peut illustrer la méthode des rangs moyens sur un petit exemple.

Exemple 2.1.1 *Considérons la variable d'intérét X correspond au niveau d'anxiété d'enfants face à la socialisation orale dans des sociétés primitives. On prend $n_1 = 15$ le groupe des enfants issus d'une société où la tradition orale. Les effets des maladies est "absent", avec celui où elle est "présente" $n_2 = 12$, comme il s'agit d'une échelle (niveau d'anxieté)*

On remarque que les observations été triées selon les valeurs de X croissantes.un numéro global sert

Explication oral des maladies	niveau d'anxiété	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	*absent*
present	8	10	9.5	n_1= 15
present	8	11	9.5	W_1=$\sum_{i=1}^{n_1} R(X_i) = 170.5$
absent	9	12	12	$\bar{r}_1$=$\frac{w_1}{n_1}$= 11.36
absent	10	13	16	
absent	10	14	16	*present*
absent	10	15	16	n_2= 12
absent	10	16	16	W_2=$\sum_{i=1}^{n_2} R(X_i) = 207.5$
present	10	17	16	$\bar{r}_2$=$\frac{w_2}{n_2}$= 17.29
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	

TAB. 2.1 – Niveau d'anxiété d'enfants face à la socialisation orale

à repérer les individus.Il correspond aux rangs bruts. Il ne tient pas comte des 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 $r'_1 = r'_2 = \frac{1+2}{2} = 1.5$, pour les observations $x_3 = x_4 = x_5 = x_6 = x_7 = 7$, nous produisons le rang $r'_3 = r'_4 = r'_5 = r'_6 = r'_7 = \frac{3+4+5+6}{5} = 5, etc...$

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

–Le groupe $n°1$ (tradition absente) :$n_1 = 15, W_1 = 170.5, \bar{r}_1 = 11.36$.

–Le groupe $n°2$ (tradition presente) :$n_2 = 12, W_2 = 207.5, \bar{r}_2 = 17.29$.

Si l'on se réfère aux rangs moyens le niveau d'anxiété semble plus faible dans le 1^{er} *groupe*, celui des enfants non informés des problèmes liés à la maladie (tradition orale =absent). Il reste à confirmer cela statistiqument.

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 convege vers la loi normale de moyenne $E(U)$ et de variance $V(U)$ (quantités données dans l'équation (2.2)).

Pour un test $H_0(F_x = F_y)$, nous pouvons donc définir la statistique centrée réduite

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{1}{12}(n_1 + n_2 + 1) n_1 n_2}} \sim N(0, 1)$$

•La région critique du test au niveau de signification α 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 2.1.1 *Ce test ne fait aucune hypothèse sur la forme de la distribution d'origine, il parmet de vérifier, pour une variable quantitative au risque d'erreur α choisi (0.05 bien souvent), si deux échantillons non appariés sont issus d'une même population, s'il est moins précis que son homologue paramétrique, il est aussi plus robuste car permet de déclarer n'importe quel type de différence entre deux échantillons.*

Remarque 2.1.2 *Ce type de test est fortement recommendé quand les échantillons sont de petites tailles. En d'autre terme on préfére le test de Mann-Whitney-Wilcoxon au test de student si l'effectif n de l'échantillon est faible et que la forme des distribitions n'est pas certaine.*

La corréction de la continuité

Lorsque les effectifs sont de taille modérée, nous pouvons améloirer l'approximation normale en introduisant la corréction de continuité, donc la statistique du test sera comme suit :

$$|Z| = \frac{|U - E(U)| - 1/2}{\sqrt{V(U)}} \tag{2.3}$$

• La règle de décision n'est pas modifiée.

Exemple 2.1.2 *Considérons un exemple de comparaison de l'indice de masse corporelle (IMC) des lycéens masculins d'une classe de terminale selon le niveau de leur activité sportive :*

$Numéroglobal$	$numérodanslegroupe$	IMC	$Sport$	$Rang$
1	1	22.8	$DAILY$	1
2	2	23.4	$DAILY$	3
3	3	23.6	$DAILY$	4
4	4	23.7	$DAILY$	5
5	5	24.8	$DAILY$	6
6	6	26.1	$DAILY$	8
7	7	30.2	$DAILY$	12
8	8	23	$NEVER$	2
9	9	26	$NEVER$	7
10	10	26.3	$NEVER$	9
11	11	27.3	$NEVER$	10
12	12	28.7	$NEVER$	11
13	13	33.5	$NEVER$	13
14	14	35.3	$NEVER$	14

TAB. 2.2 – Comparaison des IMC selon l'activité sportive

n_1	7	n_2	7
W_1	39	W_2	66
$rbarre-1$	5.571	$rbarre-2$	9.429

Sans corréction :on trouve

$U_1 = 11$, $U_2 = 38$, alors $U = 11$ et $E(U) = 24.5000$, $V(U) = 61.2500$, et la valeur de $|Z| = 1.7250$, et $z_{1-\frac{0.05}{2}} = 1.96$ avec $p = 0.0845$.

Avec corréction de contnuité :

$|Z| = 1.6611$, $p = 0.0967$

$|Z| = \frac{|11-24.5|-0.5}{\sqrt{61.2500}} = 1.6611$, $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$.

Tandi que $|Z| < z_{1-\frac{\alpha}{2}}$, nous pouvons acceptte l'hypothèse nulle, alors les données sont compatibles avec l'égalité des IMC dans les groupes.

Remarque 2.1.3 *La corréction est négligeable, lorsque les effectifs sont élevés.*

Correction pour les ex-aequo

Lorsque les ex-aequo sont associés à des individus du même groupe, la statistique du test n'est pas modifiée. En revanche, lorsque des individus de groupes différents présentent la même valeur et se voient donc attribués des rangs (moyens) identiques, la statistique du test est modifiée par rapport à la méthode des rangs aléatoires. Néanmoins la différence est négligeable et la variance de la statistique doit être corrigée. Ainsi, lorsque nous voulons utiliser l'approximation normale pour définir la région critique du test, nous devons utiliser la formule suivante :

$$\widetilde{V}(U) = V(U) \times (1 - \frac{1}{n^3 - n} \sum_{g=1}^{G} t_g(t_g^2 - 1)) \tag{2.4}$$

où $n = n_1 + n_2$ est l'éffetif total, G est le nombre de valeurs distinctes dans l'échantillon Ω et t_g est le nombre d'observation associée à la $g - ième$ valeur.

Remarque 2.1.4 *Les principaux logiciels de statistiques utilisent systématiquement la correction pour ex-aequo lors du calcul de la statistique centrée réduite pour l'approximation normal.*

Exemple 2.1.3 *Reprenons l'exemple précédent d'anaxiété des enfants. Nous complétons la feuille de calcul pour déjà construire la statistique du test. Nous y insérons les deux estimations de la variance, sans et avec la corréction pour ex-aequo :*

⋆ **Sans la corréction pour ex-aequo :** A partir de la somme des rangs, nous calculons

$$U_1 = 170.5 - \frac{15(15+1)}{2} = 50.5,\ U_2 = 207.5 - \frac{12(12+1)}{2} = 129.5.$$

Nous en déduisons $U = \min(50.5, 129.5) = 50.5$.

-L'éspérance $E(U) = \frac{15 \times 12}{2} = 90$.

-La variance $V(U) = \frac{(15+12+1) \times 15 \times 12}{12} = 420$.

-La statistique $|Z|$ s'obtient par le rapport (sans la corréction de continuité) :

$$|Z| = \frac{|50.5 - 90|}{\sqrt{420}} = 1.9274$$

La $p-value$ du test est $p = 0.00061$, au risque 5%,nous conlurons à une différence significative entre les niveaux d'anaxiété des enfants.

⋆ **Avec la corréction pour ex-aequo** :

$tableau\ des\ valeurs\ uniques$				
g	$valeur$	$Rang\ associé$	t_g	$t_g(t_g^2-1)$
1	6	1.5	2	6
2	7	5	5	120
3	8	9.5	4	60
4	9	12	1	0
5	10	16	7	336
6	11	20.5	2	6
7	12	24.5	6	210

TAB. 2.3 – Tableau des valeurs uniques Avec la corréction pour ex-aequo

En tenant compte des ex-aequo,nous devons calculer le nombre de valeurs uniques différentes G et les effetifs correspondants (t_g). C'est le rôle du tableau "Tableau des valeurs uniques". Les valeurs que l'on retrouve sont : $v_1 = 6, v_2 = 7, ..., v_7 = 12$, soit $G = 12$. Nous comptabilisons les efféctifs associes $t_1 = 2, t_2 = 5, ..., t_7 = 6$. Nous pouvons donc former la somme :

$$\sum_{g=1}^{G=7} t_g(t_g^2 - 1) = 6 + 120 + 60 + 0 + 336 + 6 + 210 = 738.$$

Sachant que $n = 15 + 12 = 27$, la nouvelle variance,en introduisant le fateur de corréction tenant compte des ex-aequo devient (en utilisant l'équation(2.4)),

$$\widetilde{V}(U) = 420 \times \left(1 - \frac{738}{27^3 - 27}\right) = 404.2307.$$

Alors, $|Z| = \frac{|50.5-90|}{\sqrt{420}} = 1.9274$ et la probabilité critique du test est $p = 0.00056$.

La correction est assez faible. Ce qui est sûr en tous les cas, c'est qu'introduire la correction pour valeurs ex-aequo ne peut que réduire la variance estimée. La statistique centrée et réduite est à l'inverse augmentée. Les résultats seront toujours un peu plus significatifs. De fait, ne pas tenir compte des exaequo correspond à favoriser l'acceptation de l'hypothèse nulle.

La variante de Wilcoxon

Historiquement, le test de **Wilcoxon (1945)** est antérieur à celui de **Mann et Whitney (1947)**. Ils sont totalement équivalents. Pour le test de Wilcoxon, nous devons choisir un groupe de référence, par convention le 1^{er}, avec $n_1 < n_2$, et si $n_1 = n_2$, il faut que $W_1 < W_2$,. La statistique de Wilcoxon W_s correspond simplement à la somme des rangs W_1. Sa variance est strictement identique à la variance de la statistique de **Mann et Whitney** $V(W_s) = V(U)$. Son espérance dépend du groupe de référence *:*

$$E(W_S) = \frac{n_1(n_1 + n_2 + 1)}{2}.$$

Pour l'approximation normale, la statistique centrée et réduite est construite exactement de la même manière. Pour un test bilatéral, nous utilisons :

$$|Z| = \frac{W_s - E(W_s)}{\sqrt{V(W_s)}}$$

Nous constaterons alors que les résultats concordent parfaitement et que les statistique peuvent se déduire l'une de l'autre :

$$U = \frac{n_1(n_1 + 2n_2 + 1)}{2} - W_s, \qquad W_s = n_1 n_2 + \frac{n(n+1)}{2} - U.$$

Une autre vision 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 : le test de **Mann-Whitney** défini par

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

On a dans ce cas $U_{n,m} = W - \frac{m(m+1)}{2}$. L'espérance et la variance de W s'écrivent :

$$E(W) = \frac{n(N+1)}{2}, \qquad var(W) = \frac{nm(N+1)}{12}$$

• La règle de décision est la même.

Exemple 2.1.4 *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 :*

échantillon femelle ($n = 6$) : $24, 30, 30, 30, 38, 40$

échantillon mâle ($m = 4$) : $20, 24, 26, 28$

On veut tester si le poids des mâles et femelles sont les mêmes ou non ?

Les hypothèses du test :

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

Pour plus de robustesse, on utilise le test de Mann-Whitney suivant la manière dont on calcul les rangs pour chaque échantillon. Notons : $X = 24, 30, 30, 30, 38, 40, Y = 20, 24, 26, 28$

Ensuite, on va classer 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 $H_0(F_X = Fy)$ **:**

$$U = \frac{U_{n,m} - \frac{nm}{2}}{\sqrt{\frac{nm(n+m)}{12}}} = \frac{25 - \frac{6*4}{2}}{\sqrt{\frac{6*4(10+1)}{12}}} = 0.0041$$

où $z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.05}{2}} = z_{0.75} = 1.96$. Comme $U < z_{1-\frac{\alpha}{2}}$, alors on acceptte H_0, (donc $F_0 = F_1$). On peut conclure que le poids des *mâles et femelles sont le mêmes.*

2.1.1 Test de Kruskal-Wallis pour $K \geq 2$

Parfois appelé **test H,** ce test ultilise les rangs de données d'échantillon de trois populations indépendantes ou plus. L'objectif de ce test est comparer la distribution d'une variable quantitative X entre K groupes indépendantes (extension à plus de 2 groupes du test de Mann-Whitney-Wilcoxon), il est utilise pour tester :

$$\begin{cases} H_0 : F_1(X) = F_2(X + \theta) = ... = F_k(X + \theta), \theta = 0 \;\; \textit{distributions identiques} \\ H_1 : \exists i \neq j \;\; | \; F_i(X) = F_j(X + \theta), \theta \neq 0 \;\; \textit{distributions diffférents} \end{cases}$$

L'édée du test est :

* Combiner tous les échantillons en un seul et affecter un rang à chaque valeur.

* Pour chaque échantillon, calculer la taille d'éhantillon et la somme des rangs.

* Calculer la statistique du test.

La statistique de test et la région critique sont comme suit :

1. Transformation en rangs (r_{ij} rang de l'observation X_{ij} parmi les n observations).

2. Calcul de la moyenne des rangs pour chaque groupe par la relation suivante :

$$\overline{w_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}.$$

3. Calcul de la moyenne globale des rangs par :

$$\overline{w} = \frac{1}{k}\sum_{j=1}^{k} \overline{w_{ij}}$$

Sous H_0 la statistique est définie de la manière suivante :

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{k} n_i(\overline{w_i} - \overline{w})^2 \sim \chi^2_{(k-1)}$$

La région critique : $H > \chi^2_{(k-1)}$ au seuil $1-\alpha$.

Remarque 2.1.5 *Approximation du χ^2 valable si $n_i > 5, \forall i \in \{1, ..., k\}$.*

• Si H_0 est vraie, alors les $\overline{w_i}$ sont proches de $\overline{w} \rightarrow H$ *tens vers* 0

• Plus H s'écarte de 0, plus on aura tendance à rejeter H_0.

On peut simplfier l'exprission ci-dessus. On retrouve plus couramment la formule suivante dans la littérature :

$$H = \frac{12}{n(n+1)}\sum_{k=1}^{K} \frac{S_k^2}{n_k} - 3(n+1)$$

où S_k est la somme des rangs des individus appartenant ou $k - ième$ groupe.

Remarque 2.1.6 *(Equivalence avec le test de **Mann et Whitney** pour $K = 2$). L'équivalence avec*

le test de Mann et Whitney est établie par la relation entre les statistiques du test pour $K = 2$*,*

$$H = \frac{12}{n(n+1)} \left(U - \frac{n+1}{2} \right)^2 .$$

De plus, le carré de la statistique centrée réduite Z *du test de* ***Wilcoxon-Mann-Whitney*** *est égal à la statistique* H *de* ***Kruskal et Wallis****. Ce dernier est bien une généralisation.*

Exemple 2.1.5 *Prenons l'exemple suivant, dont en mesure de la teneur en calcium de l'eau de 3 zones géographiques* $n_1 = 6, n_2 = 5, n_3 = 6$:

$zone1$	$zone2$	$zone3$
18	15	15
20	16	20
22	17	21
25	21	25
23	20	16
19		19

TAB. 2.4 – Mesure de la teneur en calcium de l'eau

Transformation en rangs sur n :

Valeurs	15	15	16	16	17	18	19	19	20	20	20	21	21	22	23	25	25
Rangs bruts	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Rangs finaux	1.5	1.5	3.5	3.5	5	6	7.5	7.5	10	10	10	12.5	12.5	14	15	16.5	16.5

1. Transformation en rangs sur **n** :

zone 1	(rang)	zone 2	(rang)	zone 3	(rang)
18	(6)	15	(1.5)	15	(1.5)
20	(10)	16	(3.5)	20	(10)
22	(14)	17	(5)	21	(12.5)
25	(16.5)	21	(12.5)	25	(16.5)
23	(15)	20	(10)	16	(3.5)
19	(7.5)			19	(7.5)

2. Calcul des moyennes des rangs par zones et moyenne globale des rangs :

$$\overline{w_1} = 11.5, \overline{w_2} = 6.5, \overline{w_3} = 8.58, \overline{w} = 9$$

3. Calcul de la statistique de test :

$$\begin{aligned} H &= \frac{12}{n(n+1)}\sum_{i=1}^{k} n_i(\overline{w_i} - \overline{w})^2 \\ &= \frac{12}{17(17+1)}\left[6 \times (11.5-9)^2 + 5 \times (6.5-9)^2 + 6 \times (8.58-9)^2\right] = 2.74 \end{aligned}$$

4. Région critique et conclusion : les $n_i \geq 5, 1 \leq i \leq 3$, donc $H \sim \chi^2_{(k-1)}$, le quantile de la loi du $\chi^2 = 5.99$.

Comme $2.74 < 5.99$ donc on acceptte H_0.

Alors, les teneurs en calcuim sont distribuées de la même manière parmi les 3 zones géographiques.

-Passons à la seconde formulation. Nous calculons :

$$B = \sum_{k=1}^{K} \frac{S_k}{n_k} = \frac{(69)^2}{6} + \frac{(32.5)^2}{5} + \frac{(51.5)^2}{6} = 1446.79$$

Nous en déduisons aussi :

$$\begin{aligned} H &= \frac{12}{n(n+1)} \times B - 3(n+1) \\ &= \frac{12}{17(17+1)} \times 1446.799 - 3(17+1) = 2.7368. \end{aligned}$$

Nous avaons bien la même valeur de H.

Distribution asymptotique

Lorsque les effectifs sont assez élevés, en pratique $n_k > 5\,;\, \forall k$, la distribution de H peut être approximée par une loi du χ^2 à $(K-1)$ degrés de liberté lorsque H_0 est vrai. En effet, n'oublions pas que les sommes de rangs S_k sont asymptotiquement normaux. De fait, toute statistique de la forme :

$$\sum_{k=1}^{K} \frac{[S_k - E(S_k)]^2}{V(S_k)}$$

suit une loi du χ^2à $(K-1)$ degrés de liberté lorsque compte tenu du fait que les quantités S_k sont reliées par une relation linéaire. La région critique du test au risque α s'écrit

$$R.C : H \geq \chi^2_{K-1}(1-\alpha).$$

Traitement des ex-aequo

Lorsque les données comportent des ex-aequo, nous utilisons le principe des rangs moyenes et la statistique du test devra être corrigé. Soit G le nombre de valeurs distinctes dans le fichier $(G \leq n)$. Pour la valeur $n \circ g$,nous observons t_g valeurs.La statistique ajustée s'écrit :

$$\widetilde{H} = \frac{H}{1 - \frac{1}{n^3-n}\sum_{g=1}^{G}\left(t_g^3 - t_g\right)}$$

Remarque 2.1.7 *La statistique H est calculée sur les rangs modifiés(c-à-dire avec les rangs moyens lorsqu'il y a des ax-aequo).*

Exemple 2.1.6 *Repronons notre exemple de mesure de la teneur en calcium de l'eau de 3 zones géographiques tel que $n_1 = 6, n_2 = 5, n_3 = 6$. Sous R, après l'installation du package "rank sum test", on peut utiliser la $p-value$ donnée par la fonction "$kruskal.test$" :*

$> x < -c(18, 20, 22, 25, 23, 19)$

$> y < -c(15, 16, 17, 21, 20)$

$> z < -c(15, 20, 21, 25, 16, 19)$

$> kruskal.test(list(x, y, z))$

$Kruskal - Wallisranksumtest$

$data : list(x, y, z)$

$Kruskal - Wallis\ chi - squared = 2.7675, df = 2, p - value = 0.2506.$

Intrerprétation :

On remarque que la valeur du statistique $H = 2.76$ (la même valeur qui est trouvée manuelement) ,et la p valeur est 2.25 est supérieur à la valeur de $\alpha = 0.95$. alors on ne peut pas rejetter l'hypothèse

nulle , ça veut dire on acceptte H_0 qui dit les teneurs en calcuim sont distruérs de la même manière parmi les 3 zones géographiques

2.2 Test des blocs de Wald-Wolfowity(1940)

Ce test est une d'autre test non paramétrique,qui est l'équivalent du test du nombre de séquences homogènes, il consiste à ranger par ordre croissant les observations du premièr sous-éhantillon ainsi que celles du deuxième sous-échantillon. Par définition, on dit bloc, c'est une suite des observations appartenant au même échantillon.

On considère la variable B donée par :

$$B = \textit{le nombre des blocs aléatoires.}$$

Si pour n_1(ou n_2) (la taille des sous-échantillons) est supérieur à 20, n'étant pas trop petit (au moins 10), on peut assurer que si les deux sous-échantillons indépendantes proviennent d'une population homogène, et leur disribution B est approximativement normale, avec la moyenne et la variance données par :

$$E(B) = \frac{2n_1n_2}{n_1+n_2} + 1, \qquad var(B) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

• La statistique du test est comme suit :

$$Z = \frac{B - E(B) + \frac{1}{2}}{\sqrt{var(B)}} \sim N(0,1)$$

•Les étapes d'application du test du blocs de Wald-Wolfowity sont :

1. Assurer que les sous-échantillons sont indépendantes et que $(n_1 \geq 20, n_2 \geq 20)$.
2. On agrégons les deux sous-échantillons pour n'en faire q'un seul taille comme $n = n_1 + n_2$.
3. On ordonnons ce nouvel échantillon en accordant un rang à chacune des observations.
4. Calculons la valeur de B.
5. Tester l'hypothèse nulle d'homogénéité en calculant la valeur observée de la statistique sous

l'hypothèse

$$\mathbf{Z}_{obs}=\frac{B-\frac{2n_1n_2}{n_1+n_2}+1+\frac{1}{2}}{\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}}$$

6. Décision : on rejette l'hypothèse nulle d'homogénéité avec un risque α, si : $|\mathbf{Z}_{obs}| > z_{(1-\frac{\alpha}{2})}$, où $z_{(1-\frac{\alpha}{2})}$ est le quantille d'ordre $(1-\frac{\alpha}{2})$ de la loi normale centrée réduite.

Exemple 2.2.1 *On a fait passer une épreuve à* 31 *sujets,* 14 *hommes et* 17 *femmes. Le protocole des rangs observé est le suivant :*

Hommes : $1, 2, 3, 7, 8, 9, 10, 13, 14, 15, 23, 24, 26, 27$

Femmes : $4, 5, 6, 11, 12, 16, 17, 18, 19, 20, 21, 22, 25, 28, 29, 30, 31$

Détermination du nombre du blocs :

$MMM \quad FFF \quad MMMM \quad FF \quad MMM \quad FFFFFFF \quad MM \quad F \quad MM \quad FFFF$

111 222 3333 44 555 6666666 77 8 99 0000

Ici $B = 10$.

Soit n_1 et n_2 les effectifs des deux groupes. Pour $n_1 \leq 10$ ou $n_2 \leq 10$, on utilise des tables spécialisées. Pour $n_1 > 10$ et $n_2 > 10$, on utilise l'approximation par une loi normale. Ici :

$$E(B) = 16.35, \quad Var(B) = 7.35, \quad Z_{obs} = 2.16.$$

2.3 Test du khi-deux d'homogénéité

Le test du khi-deux peut utilisé aussi dans le cadre de la comparaison 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 :

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

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 $a_{k.}$

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

Exemple 2.3.1 *Imaginons 3 populations d'étudiant dont nous étudions le taux d'admission chez* 3 *groupes d'étudiants sur lesquels* 3 *pédagogies ont été testées.*

	Admis	*ajournés*
*Pédagoie*1	51	29
*Pédagoie*2	38	12
*Pédagoie*3	86	34

–*Objectif du* **Khi-deux** : vérifier s'il y a une différence entre les 3 pédagogies.

$\begin{cases}$ Hypothèse nulle(H_0) :il n'ya pas de différence significative dans la répartition des 3 groupes étudiés

Hypothèse alternative (H_1) :au moins 1 des 3 méthodes est plus efficace que les autres

Sous R,

$> pédago1 < -c(51, 29)$

$> pédago2 < -c(38, 12)$

$> pédago3 < -c(86, 34)$

$> tableau < -matrix(c(pédago1, pédago2, pédago3), 3, 2, byrow = T)$

$> tableau$

	[, 1]	[, 2]
[1,]	51	29
[2,]	38	12

[3,] 86 34

$> chisq.test(tableau)$

$Pearson's\ Chi-squared\ test$

$data : tableau$

$X-squared = 2.504, df = 2, p-value = 0.2859$

Ainsi, la $p-value$ ici est de 0.2859 ==> La probabilité d'obtenir ainsi de telles différences de répartition entre les 3 effectifs est ainsi de 28.59%. Cela n'implique donc pas de différence particulière.

2.4 Test de Kolmogorov-Smirnov d'homogénéité

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

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

et on cherche à tester $H_0 : F_{1_n}(x) = F_{2_m}(x)$ contre $H_1 : F_{1_n}(x) \neq F_{2_m}(x)$, on utilise la distance de **K-S** d'homogénéité suivante :

$$\begin{aligned} D_{n,m} &= (\frac{1}{n} + \frac{1}{m})^{-\frac{1}{2}} \sup |F_{1_n}(t) - F_{2_m}(t)| \\ &= \sqrt{\frac{n*m}{n+m}} \sup |F_{1_n}(t) - F_{2_m}(t)| \end{aligned}$$

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

• La région critique : on rejette H_0 si $D_{n,m} > s_{ksh}$.

Nous définissons la quantilé :

$$Z = \sqrt{\frac{n*m}{n+m}} \times D$$

La probabilité critique p du test est produite en appliquant la règle suivante :

$\star$ si $0 \leq Z < 0.27, p = 1$

$\star$ si $0.27 \leq Z < 1, p = 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25}), Q = \exp(-1.233701 \times Z^{-2})$

$\star$si $1 \leq Z < 3.1, p = 2(Q - Q^4 + Q^9 - Q^{16}), Q = \exp(-2 \times Z^2)$

$\star$si $Z \geq 3.1, p = 0$

En pratique, dès que n est suffisamment grand, on utilise des approximations des lois de K-S qui reposent sur les résultats suivants :

Théorème 2.4.1 (Kolmogorov-Smirnov 1933) *Soient $X_1, ..., X_n$ un échantillon de loi continue et F_0 une fonction de répartition continue. Alors, sous H_0,*

$$\sqrt{n}D_n \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} W_0$$

où W_0 est une variable aléatoire à densité, qui ne dépend pas de F_0 à valeurs positives, de fonction de répartition

$$P(W_0 \leq t) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2t^2), pour\ tout\ t \geq 0$$

La loi de W_0 est appelée loi de **Kolmogorov-Smirnov.** C'est une loi bien connue, et tabulée. En notant $W_{1-\alpha}$ son quantile d'ordre $1-\alpha$, la région de rejet du test d'adéquation de **Kolmogorov-smirnov** de niveau asymptotique α est

$$R_{n,\alpha} = \left\{D_n \geq W_{1-\alpha}/\sqrt{n}\right\}$$

En utilsant uniquement l'approximation par le premier terme

$$\begin{aligned} P_{H_0}(D_n \geq t/\sqrt{n}) &\simeq P(W_0 \geq t) \\ &\simeq 2\exp(-2t^2), pour\ t \geq 0 \end{aligned}$$

On obtient comme valeurs approchées pour les quantiles dès que n est assez grand (typiquement

$n \geq 100$)

$$S_{n,1-\alpha} \simeq W_{1-\alpha}/\sqrt{n} \simeq \sqrt{\frac{1}{2n}\ln(\frac{2}{\alpha})} \tag{2.5}$$

Aussi, des régions de rejet de niveau asymptotique $0.01, 0.05$ *ou* 0.1 sont respectivement : $R_{n,0.01} = \{D_n \geq 1.63/\sqrt{n}\}$, $R_{n,0.05} = \{D_n \geq 1.36/\sqrt{n}\}$ et $R_{n,0.1} = \{D_n \geq 1.22/\sqrt{n}\}$.

Les régions de rejet de niveau asympotique α obtenues pas l'approximation (2.5) sont en fait, quelle que soit la taille n de l'échantillon, des régions de rejet de niveau au plus α d'aprés le résultat suivant :

Théorème 2.4.2 (Massart 1990) *Quels que soient* $n \in \mathbb{N}^*$ *et* $t \geq 0$

$$P(\sqrt{n}ks \geq t) \leq 2\exp(-2t^2).$$

Par ailleurs le théorème précedante permet également de calculer une valeur approchée de la $p-valeur$: c'est ainsi que la $p-valeur$ est calculée dans le logiciel R dès que $n \geq 100$.

Exemple 2.4.1 *On souhaite savoir est-ce que la capacité à maintenir son équilibre lorsque l'on est concentré est différente selon l'âge ? Pour répondre à cette question,* $n = 17$ *observations ont été recueillies. Des personnes ont été placées sur un plateau mouvant. Elles devaient réagir en appuyant sur un bouton lorsque des signaux arrivaient à intervalles irréguliers. Dans le même temps, elles devaient se maintenir sur le plateau. On a mesuré alors l'amplitude des corrections, d'avant en arrière, effectuées pour rester debout. Les personnes sont subdivisées en* 2 *groupes : les vieux* $(n_1 = 9)$ *et les jeunes* $(n_2 = 8)$ *selon le tablau suivant :*

$Vieux$:	X	19	30	20	10	29	25	21	24	50
$Jeune$:	Y	25	21	17	15	14	14	22	17	

TAB. 2.5 – Test de KS d'homogénéité ; Exemple

On va tester l'homogénéité des lois de X et Y

Sous R,

$> vieux < -c(19, 30, 20, 10, 29, 25, 21, 24, 50)$

$> jeune < -c(25, 21, 17, 15, 14, 14, 22, 17)$

$> ks.test(vieux, jeune, alternative = "l")$

$$Two - sample\ Kolmogorov - Smirnov\ test$$

$data : vieux\ and\ jeune$

$D\hat{} - = 0.51389, p - value = 0.1068$

$alternative\ hypothesis : the\ CDF\ of\ x\ lies\ be\ low\ that\ of\ y$

$Warning\ message :$

$In\ ks.test(vieux, jeune, alternative = "l") :$

$impossible\ de\ calculer\ la\ p - value\ exacte\ avec\ des\ ex - aequos.$

2.5 Tests de Cramér-von Mises

Test de **Cramer-von Mises** est une d'autre test que permettant de comparer des disributions. Ce test repose sur des **carrés des écarts en valeurs absolue** entre les fonctions de répartitions empiriques. En notant S_d^2 cette somme :

• La statistique du test est :

$$T = \frac{n \times m \times S_d^2}{n+m} = \frac{n \times m}{n+m} \sum_{i=1}^{n+m} [F_1(x_i) - F_2(x_i)]^2$$

avec n et m les nombres d'observations des deux groupes.

• La région critique : on rejette H_0 au niveau de signification 5% (resp 1%) si T est supérieur à 0.461 (resp 0.743).

Nous pouvons exprimer même cette statistique en passant par les rangs des observations. Soit r_{ik} le rang de l'observation $n°i$ du sous échantillon Ω_k dans l'ensemble de l'échantillon Ω. La statistique peut être calculée à l'aide de l'expression suivante :

$$T = \frac{U}{n \times m(n+m)} - \frac{4n \times m - 1}{6(n+m)}$$

où $U = n\sum_{i=1}^{n}(r_{i1} - i)^2 + m\sum_{i=1}^{m}(r_{i2} - i)^2$.

2.5.1 Traitement des ex-aequo

Lorsq'il y a deux individus ou plus prennet la même valeur (des ex-aequo), la formulation ci-dessus doit être modifiée pour tenir compte. Soit G le nombre d'observation distinctes ($G \leq n$), pour chaque valeur v_g, nous décomptons le nombre d'observations correspondantes t_g. La statistique de Cramer-von Mises s'écrit alors de la manière suivante :

$$T_{ae} = \frac{n \times m}{(n+m)^2} \sum_{g=1}^{G} t_g \times [F_1(v_i) - F_2(v_i)]^2$$

Exemple 2.5.1 *Reprenons l'exemple de comparaison de l'indice de masse corporelle (IMC) des lycéens masculins d'une classe de terminale selon le niveau de leur activité sportive : journalier="daily" ($n = 7$ élèves) ou jamais="never" ($m = 7$ élèves) :*

IMC	Sport	rang	i	Ecart	Ecart^2
26.1	DAILY	1	1	0	0
23.6	DAILY	3	2	1	1
23.4	DAILY	4	3	1	1
30.2	DAILY	5	4	1	1
22.8	DAILY	6	5	1	1
24.8	DAILY	8	6	2	4
23.7	DAILY	12	7	5	25
26	NEVER	2	1	1	1
26.3	NEVER	7	2	5	25
23	NEVER	9	3	6	36
33.5	NEVER	10	4	6	36
28.7	NEVER	11	5	6	36
27.3	NEVER	13	6	7	49
35.3	NEVER	14	7	7	49

TAB. 2.6 – Comparaison de l'IMC des lycéens masculins : Test de CM

–Les données sont organisées en groupes selon la variable Sport. Nous créons un second tableau où nous reprenons le rang des observations. Les rangs sont calculés sur la globalité de l'échantillon c.-à-d la valeur 22.8 a le rang $r_{11} = 1$ dans tout l'échantillon, la seconde valeur 23.4 possède le rang $r_{21} = 2$, etc. Nous faisons de même pour le second sous échantillon c.-à-d pour la valeur $x_{12} = 23$, nous avons $r_{12} = 2$, pour $x_{22} = 26$, nous obtenons $r_{22} = 7$, etc. Nous plaçons en face des rangs le numéro d'observation dans le sous-échantillon.

– Calculant maintenant la quantité U :

– Pour la première somme : $S_1 = (1-1)^2 + (3-2)^2 + ... + (12-7)^2 = 33$.

Pour la seconde somme : $S_2 = (2-1)^2 + (7-2)^2 + ... + (14-7)^2 = 232$

Alors

$$U = n\sum_{i=1}^{n}(r_{i1} - i)^2 + m\sum_{i=1}^{m}(r_{i2} - i)^2$$
$$= 7 \times 33 + 7 \times 232 = 1855.$$

– Nous disposons d'un résultat qui permet déjà de porter un jugement sur la signi cativité de l'écart. En effet, les tables de **Cramer-von Mises** fournissent quelques probabilités critiques pour les valeurs de la statistique (et les seuils critiques pour un niveau de risque choisi). Nous retrouvons la $p - value$ $p = 0 : 093240$. Au risque de 10%, nous rejetons l'hypothèse nulle.

Nous pouvons aussi appliquer la formule

$$T = \frac{U}{n \times m(n+m)} - \frac{4n \times m - 1}{6(n+m)}$$
$$= \frac{1855}{(7 \times 7)(7+7)} - \frac{4 \times 7 \times 7 - 1}{6(7+7)} = 0.382653.$$

Données comportant des ex aequo.

En présence d'ex aequo, les choses sont différentes. Repronons l'exemple de niveau d'anxiété d'enfants face à la socialisation orale dans des sociétés primitives.ce dernier contient d'ex aequo.

wilcox.test ne calcule plus de valeur exacte, et fait une approximation normale, en tenant compte des ex aequo. :

$> a < -c(6, 7, 7, 7, 7, 7, 8, 8, 9, 10, 10, 10, 10, 12, 12)$

$> p < -c(6, 8, 8, 10, 10, 10, 11, 11, 12, 12, 12, 12)$

$> wilcox.test(a, p, correct = FALSE)$

$Wilcoxon\ rank\ sum\ test$

résultats :

$data : a\ and\ p$

$W = 50.5, p - value = 0.04946$

$alternative\ hypothesis : true\ location\ shift\ is\ not\ equal\ to\ 0$

Par défaut, **wilcox.test** fait la correction de continuité :

$> wilcox.test(a, p)$résultats :

$W = 50.5, p - value = 0.05241.$

Conclusion

Dans cet travail nous avons présenté quelques types des tests non paramétriques qui ne dépend pas de l'hypothèse sur la distribution de l'échantillon, donc d'un champ d'application a priori plus large, et aussi sont des tests adaptés aux variables ordinales (ex : degré de satisfaction).

Au cours de ce mémoire nous avons étudié certains tests NP et leurs applications et nous sommes intéressés plus à l'homogénéité des lois.

Afin d'illustrer notre étude, de vérifier les comportements des tests d'homogénéité sur les tests non paramétriques étudié dans ce mémoire, nous avons procédé à des simulations numériques. C'est vraie dans ce mémoire, on a abordé l'évaluation de l'efficacité des tests NP mais cela ne veut pas dire que ce n'est pas sans inconvénients car lorsque les conditions d'application sont vérifiées : les tests NP sont moins puissants que les tests paramétriques, aussi il y a une difficulté d'interprétation : on ne compare plus des paramètres (moyenne, proportion, variance,...).

Ce travail de recherche n'est qu'une tentative étudié d'une manière objective et scientifique. L'une des perspectives de recherche que l'on peut ajouter prochainement est l'application (avec comparaison) des tests d'homogénéité sur des données réelles de notre environnement.

Bibliographie

[1] Akakpo, N. (2017). Tests statistiques. Notes de cours issues du module 4M018 Statistique Appliquée. Polycopié disponible sur la page web de l'auteur.

[2] Bettayeb, H. (2017) Test de Wilcoxon Mann-Whitney, Mémoire Master en statistique. Université de Biskra.

[3] Chesneau, C. (2016). Introduction aux tests statistiques avec R.

[4] Der Megreditchian, G. (1986). Un test non paramétrique unilatéral de rupture d'homogénéité de «K» échantillons. Revue de statistique appliquée, 34(1), 45-60.

[5] Morice, E. (1956). Quelques tests non paramétriques. Revue de statistique appliquée, 4(4), 75-107.

[6] Morice, E. (1958). Quelques tests non paramétriques. Journal de la société française de statistique, 99, 254-284.

[7] Ondo, J. C., Ouarda, T. B., Bobée, B. (1997). Revue bibliographique des tests d'homogénéité et d'indépendance (No. R500). INRS-Eau.

[8] Pizzut, A., Galle, V., Alain, L. Analyse d'homogénéité sur le cercle unité de R 2.

[9] Rakotomalala, R. (2008). Comparaison de populations. Tests Non Paramétriques..

[10] Rakotomalala, R. (2011). Pratique de la regression lineaire multiple. Diagnostic et selection de variables.

[11] Rahal A. (2014). Tests de normalité : Simulation en Logiciel R. Mémoire de master en statistique. Université de Biskra.

[12] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

Annexe A : Logiciel *R*

• Le langage **R** est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• **R** a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

• L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

Ω	:	l'ensemble fondamantale.
P	:	probabilité.
H_0	:	l'hypothèse nulle.
H_1	:	l'hypothèse alternative.
α	:	le risque de première espèse.
β	:	le risque de deuxième espèse.
$p - value$	:	la p-valeur.
$F(x)$	:	la fonction de répartition.

$F_n(x)$	:	la fonction de répartition empirique.
$x_{(i)}$	:	la statistique d'ordre associe l'éhantillon.
$\overline{x}$	:	la moyenne empirique.
s_x	:	l'écart type empirique.
$E(.)$	:	espérence mathématique.
$V(.)$	:	variance mathématique.
dll	:	degré de liberté.
$\mathbb{R}$	:	l'ensemble des nombres réelles.
θ	:	un paramètre inconnue.
w	:	la région de refuser H_0.
$\overline{w}$	:	la région d'accepeter H_0.
$K-S$	:	le test de Kolmogorov-Smirnov.
$D(P,P_0)$	:	la distance de K-S entre P et P_0.
$d_{n,\alpha}$	:	le quantile théorique du table k-s.
$khi-2$	:	le test de pearson.
$Exp(\lambda)$	:	la oi de exponentiel de paramètre λ.
$z_{1-\frac{\alpha}{2}}$	:	le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centré réduite.
λ_n^2	:	loi de khi-deux à n degré de liberté.
$\lambda^2(n-1)$	:	loi de khi-deux à $(n-1)$ degré de liberté.

Annexe C :Tables Statistiques

Test de Kolmogorov-Smirnov

Table de Kolmogorov-Smirnov pour le test d'homogénéité de 2 échantillons : valeurs critiques de D_{n_1,n_2} pour n_1 ; $n_2 \leq 20$ avec $n_1 \neq n_2$

n_1 n_2	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5		0.8	0.8	0.75	0.778	0.8	0.709	0.717	0.692	0.657	0.733	0.8	0.647	0.667	0.642	0.65
6	1		0.714	0.708	0.722	0.667	0.652	0.667	0.667	0.643	0.533	0.625	0.667	0.667	0.614	0.6
7	1	0.857		0.714	0.667	0.657	0.623	0.631	0.615	0.643	0.59	0.571	0.571	0.571	0.571	0.564
8	0.875	0.833	0.857		**0.639**	0.6	0.602	0.625	0.596	0.571	0.558	0.625	0.566	0.611	0.539	0.55
9	0.889	0.883	0.778	0.764		0.589	0.596	0.583	0.556	0.556	0.556	0.542	0.536	0.556	0.52	0.517
10	0.9	0.8	0.757	0.75	0.7		0.545	0.55	0.569	0.529	0.533	0.525	0.524	0.511	0.495	0.55
11	0.818	0.818	0.766	0.727	0.707	0.7		0.545	0.524	0.532	0.509	0.506	0.497	0.49	0.488	0.486
12	0.833	0.833	0.714	0.708	0.694	0.667	0.652		0.519	0.512	0.517	0.5	0.49	0.5	0.474	0.483
13	0.8	0.769	0.714	0.692	0.667	0.646	0.636	0.608		0.489	0.492	0.486	0.475	0.47	0.462	0.462
14	0.8	0.762	0.786	0.679	0.667	0.643	0.623	0.619	0.571		0.467	0.473	0.467	0.46	0.455	0.45
15	0.8	0.767	0.714	0.675	0.667	0.667	0.618	0.6	0.59	0.586		0.475	0.455	0.456	0.446	0.45
16	0.8	0.75	0.688	0.688	0.653	0.625	0.602	0.604	0.582	0.563	0.554		0.456	0.444	0.437	0.437
17	0.8	0.716	0.706	0.647	0.647	0.624	0.588	0.583	0.576	0.563	0.557	0.526		0.435	0.437	0.429
18	0.788	0.778	0.69	0.653	0.667	0.6	0.596	0.583	0.585	0.556	0.544	0.535	0.536		0.415	0.422
19	0.747	0.728	0.684	0.645	0.626	0.595	0.584	0.57	0.559	0.556	0.533	0.526	0.514	0.515		0.421
20	0.8	0.733	0.664	0.65	0.617	0.65	0.577	0.583	0.55	0.543	0.533	0.525	0.515	0.506	0.492	

FIG. 2.2 – Table des valeurs critiques de Dn1 ;n2 - Test de Kolmogorov Smirnov

Table de la loi normale standard

La table donne les valeurs de la fonction de répartition F_z de la loi normale $N(0,1)$.Rappelons que

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

FIG. 2.3 – -Table de la loi normale centré réduite

Table de Mann et Whitney

Valeurs critiques U crit à comparer avec la valeur observée U à partir de vos deux échantillons pour un test bilatéral.

NB : n et m représentent le nombre d'observations dans chaque échantillon.

n	p	m=2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	.001	3	3	3	3	3	_	3	3	3	3	3	3	3	3	3	3	3	3	3
	.005	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4
	.01	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	5	5
	.025	3	3	3	3	3	4	4	4	5	5	5	5	5	5	5	6	6	6	6
	.05	3	3	3	4	4	4	5	5	5	5	6	6	7	7	7	7	8	8	8
	.10	3	4	4	5	5	6	6	7	7	8	8	8	9	9	9	10	10	11	11
3	.001	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	7	7	7	7
	.005	6	6	6	6	6	6	7	7	7	7	8	8	8	9	9	9	9	10	10
	.01	6	6	6	6	6	7	7	8	8	8	9	9	9	10	10	11	11	11	12
	.025	6	6	6	7	8	8	9	9	10	10	11	11	12	12	13	13	14	14	15
	.05	6	7	7	8	9	9	10	11	11	12	12	13	14	14	15	16	16	17	18
	.10	7	8	8	9	10	11	12	12	13	14	15	16	17	17	18	19	20	21	22
4	.001	10	10	10	10	10	10	10	10	11	11	11	12	12	12	13	13	14	14	14
	.005	10	10	10	10	11	11	12	12	13	13	14	14	15	16	16	17	17	18	19
	.01	10	10	10	11	12	12	13	14	14	15	16	16	17	18	18	19	20	20	21
	.025	10	10	11	12	13	14	15	15	16	17	18	19	20	21	22	22	23	24	25
	.05	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	28	29
	.10	11	12	14	15	16	17	18	20	21	22	23	24	26	27	28	29	31	32	33
5	.001	15	15	15	15	15	15	16	17	17	18	18	19	19	20	21	21	22	23	23
	.005	15	15	15	16	17	17	18	19	20	21	22	23	23	24	25	26	27	28	29
	.01	15	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
	.025	15	16	17	18	19	21	22	23	24	25	27	28	29	30	31	33	34	35	36
	.05	16	17	18	20	21	22	24	25	27	28	29	31	32	34	35	36	38	39	41
	.10	17	18	20	21	23	24	26	28	29	31	32	34	36	38	39	41	43	44	46
6	.001	21	21	21	21	21	21	23	24	25	26	26	27	28	29	30	31	32	33	34
	.005	21	21	22	23	24	25	26	27	28	29	31	32	33	34	35	37	38	39	40
	.01	21	21	23	24	25	26	28	29	30	31	33	34	35	37	38	40	41	42	44
	.025	21	23	24	25	27	28	30	32	33	35	36	38	39	41	43	44	46	47	49
	.05	22	24	25	27	29	30	32	34	36	38	39	41	43	45	47	48	50	52	54
	.10	23	25	27	29	31	33	35	37	39	41	43	45	47	49	51	53	56	58	60
7	.001	28	28	28	28	29	30	31	32	34	35	36	37	38	39	40	42	43	44	45
	.005	28	28	29	30	32	33	35	36	38	39	41	42	44	45	47	48	50	51	53
	.01	28	29	30	32	33	35	36	38	40	41	43	45	46	48	50	52	53	55	57
	.025	28	30	32	34	35	37	39	41	43	45	47	49	51	53	55	57	59	61	63
	.05	29	31	33	35	37	40	42	44	46	48	50	53	55	57	59	62	64	66	68
	.10	30	33	35	37	40	42	45	47	50	52	55	57	60	62	65	67	70	72	75
8	.001	36	36	36	37	38	39	41	42	43	45	46	48	49	51	52	54	55	57	58
	.005	36	36	38	39	41	43	44	46	48	50	52	54	55	57	59	61	63	65	67
	.01	36	37	39	41	43	44	46	48	50	52	54	56	59	61	63	65	67	69	71
	.025	37	39	41	43	45	47	50	52	54	56	59	61	63	66	68	71	73	75	78
	.05	38	40	42	45	47	50	52	55	57	60	63	65	68	70	73	76	78	81	84
	.10	39	42	44	47	50	53	56	59	61	64	67	70	73	76	79	82	85	88	91
9	.001	45	45	45	47	48	49	51	53	54	56	58	60	61	63	65	67	69	71	72
	.005	45	46	47	49	51	53	55	57	59	62	64	66	68	70	73	75	77	79	82
	.01	45	47	49	51	53	55	57	60	62	64	67	69	72	74	77	79	82	84	86
	.025	46	48	50	53	56	58	61	63	66	69	72	74	77	80	83	85	88	91	94
	.05	47	50	52	55	58	61	64	67	70	73	76	79	82	85	88	91	94	97	100
	.10	48	51	55	58	61	64	68	71	74	77	81	84	87	91	94	98	101	104	108
10	.001	55	55	56	57	59	61	62	64	66	68	70	73	75	77	79	81	83	85	88
	.005	55	56	58	60	62	65	67	69	72	74	77	80	82	85	87	90	93	95	98
	.01	55	57	59	62	64	67	69	72	75	78	80	83	86	89	92	94	97	100	103
	.025	56	59	61	64	67	70	73	76	79	82	85	89	92	95	98	101	104	108	111
	.05	57	60	63	67	70	73	76	80	83	87	90	93	97	100	104	107	111	114	118
	.10	59	62	66	69	73	77	80	84	88	92	95	99	103	107	110	114	118	122	126

FIG. 2.4 – -Table de Wilcoxon Mann-Whitney (n=2 jusqu'a 10).

Table de Kruskal et Wallis

Cette table est extraite de l'article originel l'article est accessible via le lien sur Wikipedia http ://en.wikipedia.o

Wallis_one-way_

analysis_of_variance).

n	p	m=2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
11	.001	66	66	67	69	71	73	75	77	79	82	84	87	89	91	94	96	99	101	104
	.005	66	67	69	72	74	77	80	83	85	88	91	94	97	100	103	106	109	112	115
	.01	66	68	71	74	76	79	82	85	89	92	95	98	101	104	108	111	114	117	120
	.025	67	70	73	76	80	83	86	90	93	97	100	104	107	111	114	118	122	125	129
	.05	68	72	75	79	83	86	90	94	98	101	105	109	113	117	121	124	128	132	136
	.10	70	74	78	82	86	90	94	98	103	107	111	115	119	124	128	132	136	140	145
12	.001	78	78	79	81	83	86	88	91	93	96	98	102	104	106	110	113	116	118	121
	.005	78	80	82	85	88	91	94	97	100	103	106	110	113	116	120	123	126	130	133
	.01	78	81	84	87	90	93	96	100	103	107	110	114	117	121	125	128	132	135	139
	.025	80	83	86	90	93	97	101	105	108	112	116	120	124	128	132	136	140	144	148
	.05	81	84	88	92	96	100	105	109	111	117	121	126	130	134	139	143	147	151	156
	.10	83	87	91	96	100	105	109	114	118	123	128	132	137	142	146	151	156	160	165
13	.001	91	91	93	95	97	100	103	106	109	112	115	118	121	124	127	130	134	137	140
	.005	91	93	95	99	102	105	109	112	116	119	123	126	130	134	137	141	145	149	152
	.01	92	94	97	101	104	108	112	115	119	123	127	131	135	139	143	147	151	155	159
	.025	93	96	100	104	108	112	116	120	125	129	133	137	142	146	151	155	159	164	168
	.05	94	98	102	107	111	116	120	125	129	134	139	143	148	153	157	162	167	172	176
	.10	96	101	105	110	115	120	125	130	135	140	145	150	155	160	166	171	176	181	186
14	.001	105	105	107	109	112	115	118	121	125	128	131	135	138	142	145	149	152	156	160
	.005	105	107	110	113	117	121	124	128	132	136	140	144	148	152	156	160	164	169	173
	.01	106	108	112	116	119	123	128	132	136	140	144	149	153	157	162	166	171	175	179
	.025	107	111	115	119	123	128	132	137	142	146	151	156	161	165	170	175	180	184	189
	.05	109	113	117	122	127	132	137	142	147	152	157	162	167	172	177	183	188	193	198
	.10	110	116	121	126	131	137	142	147	153	158	164	169	175	180	186	191	197	203	208
15	.001	120	120	122	125	128	133	135	138	142	145	149	153	157	161	164	168	172	176	180
	.005	120	123	126	129	133	137	141	145	150	154	158	163	167	172	176	181	185	190	194
	.01	121	124	128	132	136	140	145	149	154	158	163	168	172	177	182	187	191	196	201
	.025	122	126	131	135	140	145	150	155	160	165	170	175	180	185	191	196	201	206	211
	.05	124	128	133	139	144	149	154	160	165	171	176	182	187	193	198	204	209	215	221
	.10	126	131	137	143	148	154	160	166	172	178	184	189	195	201	207	213	219	225	231
16	.001	136	136	139	142	145	148	152	156	160	164	168	172	176	180	185	189	193	197	202
	.005	136	139	142	146	150	155	159	164	168	173	178	182	187	192	197	202	207	211	216
	.01	137	140	144	148	153	158	163	168	173	178	183	188	193	198	203	208	213	219	224
	.025	138	143	148	152	158	163	168	174	179	184	190	196	201	207	212	218	223	229	235
	.05	140	145	151	156	162	167	173	179	185	191	197	202	208	214	220	226	232	238	244
	.10	142	148	154	160	166	173	179	185	191	198	204	211	217	223	230	236	243	249	256
17	.001	153	154	156	159	163	167	171	175	179	183	188	192	197	201	206	211	215	220	224
	.005	153	156	160	164	169	173	178	183	188	193	198	203	208	214	219	224	229	235	240
	.01	154	158	162	167	172	177	182	187	192	198	203	209	214	220	225	231	236	242	247
	.025	156	160	165	171	176	182	188	193	199	205	211	217	223	229	235	241	247	253	259
	.05	157	163	169	174	180	187	193	199	205	211	218	224	231	237	243	250	256	263	269
	.10	160	166	172	179	185	192	199	206	212	219	226	233	239	246	253	260	267	274	281
18	.001	171	172	175	178	182	186	190	195	199	204	209	214	218	223	228	233	238	243	248
	.005	171	174	178	183	188	193	198	203	209	214	219	225	230	236	242	247	253	259	264
	.01	172	176	181	186	191	196	202	208	213	219	225	231	237	242	248	254	260	265	272
	.025	174	179	184	190	196	202	208	214	220	227	233	239	246	252	258	265	271	278	284
	.05	176	181	188	194	200	207	213	220	227	233	240	247	254	260	267	274	281	288	295
	.10	178	185	192	199	206	213	220	227	234	241	249	256	263	270	278	285	292	300	307
19	.001	190	191	194	198	202	206	211	216	220	225	231	236	241	246	251	257	262	268	273
	.005	191	194	198	203	208	213	219	224	230	236	242	248	254	260	265	272	278	284	290
	.01	192	195	200	206	211	217	223	229	235	241	247	254	260	266	273	279	285	292	298
	.025	193	198	204	210	216	223	229	236	243	249	256	263	269	276	283	290	297	304	310
	.05	195	201	208	214	221	228	235	242	249	256	263	271	278	285	292	300	307	314	321
	.10	198	205	212	219	227	234	242	249	257	264	272	280	288	295	303	311	319	326	334

FIG. 2.5 – Table de Wilcoxon Mann-Whitney (n=11 jusqu'a 19)

n	p	m=2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
20	.001	210	211	214	218	223	227	232	237	243	248	253	259	265	270	276
	.005	211	214	219	224	229	235	241	247	253	259	265	271	278	284	290
	.01	212	216	221	227	233	239	245	251	258	264	271	278	284	291	298
	.025	213	219	225	231	238	245	251	259	266	273	280	287	294	301	309
	.05	215	222	229	236	243	250	258	265	273	280	288	295	303	311	318
	.10	218	226	233	241	249	257	265	273	281	289	297	305	313	321	330

FIG. 2.6 – Table de Wilcoxon Mann-Whitney n=20

USE OF RANKS IN ONE-CRITERION VARIANCE ANALYSIS 617

TABLE 6.1 (Continued)

Sample Sizes			H	True Probability	Approximate minus true probability		
n_1	n_2	n_3			χ^2	Γ (Linear Interp.)	B (Normal Interp.)
5	4	3	7.4449	.010	+.014	+.004	−.004
			7.3949	.011	+.014	+.004	−.004
			5.6564	.049	+.010	−.005	−.004
			5.6308	.050	+.010	−.006	−.004
			4.5487	.099	+.004	−.013	+.003
			4.5231	.103	+.001	−.016	−.000
5	4	4	7.7604	.009	+.011	+.003	−.002
			7.7440	.011	+.010	+.002	−.003
			5.6571	.049	+.010	−.004	+.000
			5.6176	.050	+.010	−.004	+.001
			4.6187	.100	−.001	−.016	+.003
			4.5527	.102	+.001	−.014	+.005
5	5	1	7.3091	.009	+.016	−.002	−.009
			6.8364	.011	+.022	+.001	−.009
			5.1273	.046	+.031	−.003	−.005
			4.9091	.053	+.032	−.002	−.002
			4.1091	.086	+.042	+.007	+.020
			4.0364	.105	+.028	−.007	+.008
5	5	2	7.3385	.010	+.016	+.004	−.004
			7.2692	.010	+.016	+.004	−.004
			5.3385	.047	+.022	+.003	+.006
			5.2462	.051	+.022	+.002	+.007
			4.6231	.097	+.002	−.018	−.005
			4.5077	.100	+.005	−.016	−.001
5	5	3	7.5780	.010	+.013	+.004	−.001
			7.5429	.010	+.013	+.004	−.002
			5.7055	.046	+.012	−.003	+.000
			5.6264	.051	+.009	−.005	−.002
			4.5451	.100	+.003	−.012	+.007
			4.5363	.102	+.002	−.014	+.005
5	5	4	7.8229	.010	+.010	+.003	−.002
			7.7914	.010	+.010	+.003	−.002
			5.6657	.049	+.010	−.003	+.001
			5.6429	.050	+.009	−.003	+.001
			4.5229	.099	+.005	−.009	+.010
			4.5200	.101	+.004	−.010	+.008
5	5	5	8.0000	.009	+.009	+.003	−.002
			7.9800	.010	+.008	+.002	−.003
			5.7800	.049	+.007	−.005	−.001
			5.6600	.051	+.008	−.004	+.001
			4.5600	.100	+.003	−.010	+.008
			4.5000	.102	+.004	−.009	+.009

FIG. 2.7 – Table des probabilités critiques pour le test de Kruskal et Wallis

Test de Cramer - von Mises

Table de Cramer - von Mises au niveau de signi cation 10% pour les petits e ectifs ($n_1, n_2 \leq 7$).Reproduite à partir de http ://projecteuclid.org/DPubS/Repository/1.0/Disseminate ?view=body&id=pdf

N	M	u	$\Pr\{U \geqq u\}$	t	Normalized t
4	6	472	$\frac{18}{210} = .085\ 714$	.383 333	
		468	$\frac{22}{210} = .104\ 762$	.366 667	
4	7	634	$\frac{32}{330} = .096\ 970$	.376 623	
		631	$\frac{34}{330} = .103\ 030$	.366 883	
5	6	718	$\frac{46}{462} = .099\ 567$	.372 727	
		710	$\frac{48}{462} = .103\ 896$	.348 485	
5	7	967	$\frac{78}{792} = .098\ 485$	.371 825	
		963	$\frac{80}{792} = .101\ 010$	.362 302	
6	6	1020	$\frac{43}{462} = .093\ 074$	.375 000	.371 314
		1008	$\frac{59}{462} = .127\ 706$	.347 222	.342 078
6	7	1374	$\frac{166}{1716} = .096\ 737$	.375 458	.372 127
		1373	$\frac{172}{1716} = .100\ 233$	.373 626	.370 207
		⋮			
		1362	$\frac{194}{1716} = .113\ 054$	.353 480	.349 084
		1359	$\frac{196}{1716} = .114\ 219$	.347 985	.343 324
7	7	1855	$\frac{160}{1716} = .093\ 240$	.382 653	.379 626
		1841	$\frac{185}{1716} = .107\ 809$	.362 245	.358 330
		1827	$\frac{197}{1716} = .114\ 802$	.341 837	.337 034
∞	∞		.10	.347 30	.347 30

FIG. 2.8 – -Table des valeurs critiques à 10% pour la statistique de Cramer - von Mises

Table de la loi de Khi deux

X étant une variable aléatoire de loi du χ^2 à n degrés de libertés et p un réel de $[0; 1]$; la table donne la valeur de $z_{n,p} = F^{-1}_{\chi^2_n}(1-p)$ telle que $P(X \succ z_{n,p}) = p$.

ν	P = 0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,00004	0,0002	0,001	0,0039	0,0158	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,39	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672

FIG. 2.9 – -Table de la loi de khi-deux

ملخص

نقدم في هذه المذكرة لمحة حول الاختبارات اللاوسيطية. نذكر ببعض المفاهيم الأساسية ونعرض بشكل عام بعض الاختبارات التي تقارن دالة التوزيع التجريبية بدالة التوزيع النظرية. وبشكل خاص نتطرق إلى اختبارات التجانس بين توزيعين أو أكثر.

" نقدم بعض التطبيقات للنتائج النظرية المطروحة باستخدام معطيات محاكاة. Rوبالاعتماد على البرنامج الإحصائي "

Résumé

Nous présentons dans ce mémoire un aperçu sur les tests non paramétriques. Nous rappelons quelques notions et concepts fondamentales, nous exposons en générale certains tests qui compare la fonction de distribution empirique à la fonction de répartition, ensuite on va aborder au test d'homogénéité entre deux distributions ou plus. A l'aide du logiciel d'analyse statistique 'R', nous proposons quelques applications permettant d'illustrer les résultats théoriques obtenus, sur des données simulées.

Abstract

We present in this memory a preview about the non-parametric-tests. We recall some fundamental notions and concepts, we present in general some tests which compare the empirical distribution function to the distribution function, then we will approach the homogeneity test between two or more distributions. Using the statistical analysis software 'R', we propose some applications allowing to illustrate the theoretical results obtained, on simulated data.