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

HIRECHE Sara

Titre :

# Sur quelques tests non-paramétriques : théorie et application

Membres du Comité d'Examen :

| | | | |
|---|---|---|---|
| Dr. | BENELMIR Imene | UMKB | Président |
| Pr. | NECIR Abdelhakim | UMKB | Encadreur |
| Dr. | CHINE Amel | UMKB | Examinateur |

juin 2021

## Dédicace

Je dédie ce modeste travail à :

Ma mère qui m'a initié à la vie, qui m'appris la modestie,

L'esprit de mon défunt Père,

Mon Mari : Mohamed Taha Traka,

Mes belles sœurs : Basma, Zahra, Ladmia, Bothaina,

Mes chères frères : Amar, Saad Amir,

Ma chère Tante Noura,

Tout la famille : Hireche, Bourdji, Traka,

Tout mes amies,

et toute la promotion "Master" Statistique 2021.

## REMERCIEMENTS

" *Yesterday is history, tomorrow is a mystery, today is a gift of God which is why we call it the present* (Bill Keane)"

*Tout d'abord, nous remercions "ALLAH" le tout Puissant de nous avoir donné le courage, la volonté et patience de mener à terme ce présent travail..*

Un grand merci au Professeur Necir Abdelhakim, le directeur de mon mémoire, qui a dirigé et suivi mes travaux

et qui a toujours su me faire confiance et m'apporter l'aide nécessaire, tant sur le plan scientifique que moral.

Nous sincères remerciements vont également à tous les enseignants du Département de Mathématiques de l'Université Mohamed Khider Biskra, ainsi qu'à tous les gens de près ou de loin qui ont contribué à la réalisation de ce modeste travail.

Merci infiniment à nos familles et nos amies pour leur soutien et leurs encouragements.

# Table des matières

# Table des figures

# Liste des tableaux

# Introduction

En statistique, un test est une technique et des méthodes qui permettent d'analyser les données issues de l'observation et de permet de trancher entre deux hypothèses à la vue des résultats d'un échantillon, en quantifient le risque associé à la décision prise. Les tests statistiques jouent un rôle efficace et important dans l'analyse des données collectées afin de les présenter avec précision. Les tests statistiques visent à atteindre les résultats corrects qui ont les résultats à généraliser au niveau sociétal. Il existe deux types de test : Les tests statistiques dont l'analyse statistique dépend de la distribution normale dans la recherche scientifique s'appelle la statistique paramétrique et les tests non paramétriques sont appelés distribution libre et font référence à l'utilisation de tests statistiques qui ne font pas d'hypothèses sur la distribution des erreurs. Ces tests sont moins puissants que les tests paramétriques, et la stratégie alternative consiste à utiliser des tests non paramétriques ou non paramétriques, comme on les appelle, en convertissant les données en une distribution normale ou proche de celle-ci.

Si le chercheur souhaite choisir le test statistique approprié, il doit considérer les éléments suivants :

- ***Question de recherche :*** le chercheur doit se demander si la question de recherche principale concerne la relation, ou la prédiction entre mesures, ou la comparaison entre groupes.
- ***Conception de la recherche :*** Combien de groupes l'étude comprendra-t-elle et

unit-elle une relation entre ces groupes ? Existe-t-il au moins deux groupes liés ou indépendants ?

– ***Distribution des données*** :La distribution des variables importantes est-elle discrète ou continue ?

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

**Le premier chapitre** *:* Généralités sur les tests statistiques, dans ce chapitre nous énonçons (ou rappelons) un certain nombre de généralités autour des tests d'hypothèses. Nous présentons alors toutes les notions principales qui dépendent d'un test comme hypothèse, erreur, risque, région d'acceptation et de rejet,..., de plus on montre la démarche d'un test qui nous conduit à prendre une décision concernant l'hypothèse posée, et énonçons aussi des propriétés fondamentales sur quelques tests regroupés sur deux test : Test basé sur les distributions, comme le test de Kolmogorov-Smirnov sur un seul échantillon, le test de Lilliefors, le test de Anderson-Darling, le test de Cramer-Von Mises, le test de Shapiro-Wilk, le test de Khi-deux, le test de Kuiper et Test basé sur les rangs, comme le test de Wilxcone, le test de Mann-Whiteny, le test de la Médiane.

**Le deuxiéme Chapitre** *:* Nous intéressons sur les tests d'homogénéités et d'indépendance entre deux distributions, comme le test de Kruskal-Wallis, test de Khi-deux d'homogénéité et d'indépendance, test de Kolmogrov-Smirnov.

Dans ce mémoire, on utilisera les logiciels R et RStudio pour traiter plusieurs exemples d'application numérique.

# Chapitre 1

# Généralité sur les Tests Statistiques paramétrique

## 1.1 Principe d'un test

Un test statistique est une procédure statistique de décision (rejeter ou accepter une hypothèse) à partir de l'étude d'un ou plusieurs échantillons aléatoire. Ils sont des outils statistiques d'aide de la décision. Ils vont permettre de comparer un où plusieurs échantillon, et de valider ou d'invalider une hypothèse donnée.

## 1.2 Elément de test

Un test des hypothèses est une règle de décision qui en présence de deux hypothèses $H_0$ et $H_1$ sur la base des données observées et avec des risques d'erreur déterminés, d'accepter ou de refuser une hypothèse statistique.

### 1.2.1 Hypothèse

Dans le contexte d'un test paramétrique, une hypothèse, notée $H$, est une affirmation concernant la valeur d'un paramètre (une moyenne $\mu$, une variance $\sigma^2$, etc.), on suppose on général que la distribution de la variable étudie $X$ dépend d'un paramètre $\theta$. Les deux hypothèses à tester, notées par $H_0$ et $H_1$, sont respectivement appellées **hypothèse nulle** et **hypothèse alternative :**

- L'hypothèse nulle est $H_0 : \theta = \theta_0$ (où $\theta_0$ est une valeur comme donnée).
- L'hypothèse alternative est $H_1$, selon le problème considéré, a les trois formes suivantes :

$$H_1 : \theta \neq \theta_0 \text{ (une hypothèse dite bilatérale),}$$

$$H_1 : \theta < \theta_0 \text{ (une hypothèse dite unilatéral à gauche),}$$

$$H_1 : \theta > \theta_0 \text{(une hypothèse dite unilatérale à droite).}$$

On note que les hypothèses à tester sont de la forme générale suivante :

$H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$, où $\Theta_0$ et $\Theta_1$ sont telles que $\Theta_1 \subset \Theta, \Theta_2 \subset \Theta$ et $\Theta_0 \cap \Theta_1 = \phi$.

Dans le cas d'un test paramétrique on distingue deux types d'hypothèses : hypothèse simple et hypothèse multiple.

- Une hypothèse $H$ est dite simple si elle a la forme "$\theta = \theta_0$" où $\theta_0 \in \Theta$, avec $\Theta$ désignant l'ensemble de toutes les valeurs de $\theta$.
- Une hypothèse $H$ est dite multiple si elle a la forme "$\theta \in \theta_0$" où $\theta_0 \subset \Theta$ ayant deux éléments ou plus.

### 1.2.2 Erreurs

La décision d'un test se base sur les données d'un échantillon aléatoire de la population. Il y a donc quatre cas possibles :

– Accepter $H_0$ et elle est vraie(rejeter $H_1$).

– Rejeter $H_0$ et elle est fausse (accepter $H_1$).

– Rejeter $H_0$ et elle est vraie (accepter $H_1$).

– Accepter $H_0$ et elle est fausse (rejeter $H_1$).

La décision prise est bonne dans les deux premiers cas, mais elle est erronée dans les deux derniers. Le tableau suivant donne un résumé sur les erreurs possibles.

| | Réalité | |
|---|---|---|
| Décision | $H_0$ vrai | $H_1$ fausse |
| Accepter $H_0$ | Pas d'erreur | Erreur de deuxième espèce |
| Rejeter $H_0$(accepter $H_1$) | Erreur de première espèce | Pas d'erreur |

TAB. 1.1 – Tableau des erreurs possibles dans un test

– **Erreur de première espèce** : c'est l'erreur qui consiste à rejeter l'hypothèse $H_0$ alors qu'elle est vrai, c'est-à-dire la probabilité d'avoir un faux-positif.

– **Erreur de deuxième espèce** : C'est l'erreur qui consiste à accepter $H_0$ alors qu'elle est fausse, c'est-à-dire fausse,c'est-à-dire la probabilité d'avoir un faux-négatif.

### 1.2.3 Risques

**Définition 1.2.1** *On appelle risque, la probabilité de commettre une erreur, est aussi appelé en bref risque. On a deux types de risques :*

1. ***Risque de première espèce** : Probabilité de rejeter $H_0$ et d'accepter $H_1$ alors*

*que $H_0$ est vraie, notée $\alpha$ :*

$$\alpha = P \text{ (rejeter } H_0/H_0 \text{ est vrai)},$$

*la valeur maximal de risque de première espèce est appelée niveau (ou seuil) critique du test.*

2. ***Risque de deuxième espèce** : Probabilité de rejeter $H_1$ et d'accepter $H_0$ alors que $H_1$ est vraie, notée :*

$$\beta = P(\text{accepter } H_0/H_0 \text{ est fausse}).$$

*On résume les quatre situations possibles dans le tableau suivant :*

| Décision | $H_0$ | $H_1$ |
|---|---|---|
| $H_0$ | $1-\alpha$ | $\beta$ |
| $H_1$ | $\alpha$ | $1-\beta$ |

TAB. 1.2 – Tableau de décision

**Remarque 1.2.1** *Les seuils critiques les plus utilisés sont $\alpha = 0.10, \alpha = 0.05$ et $\alpha = 0.01$.*

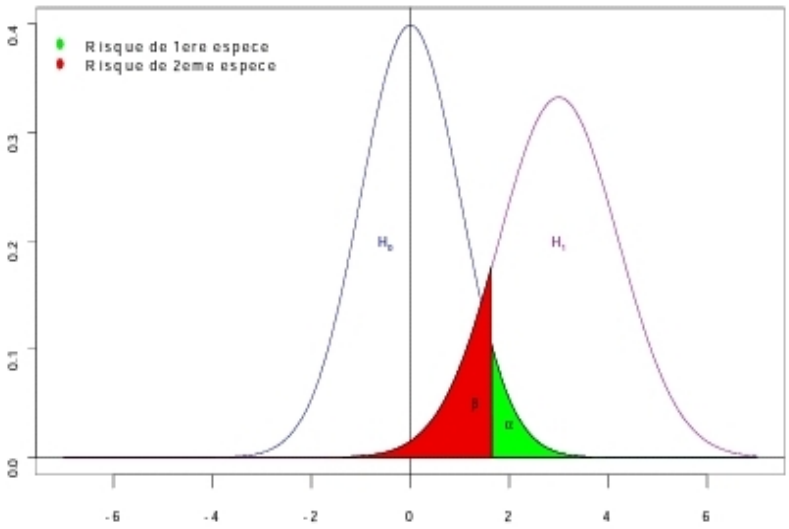


FIG. 1.1 – Risques de première et deuxième espèce

### 1.2.4 Région de rejet-d'acceptation

La région de rejet d'un test est l'ensemble des observations $(x_1, ...., x_n)$ dans $\mathbb{R}^n$ pour la quelle l'hypothèse nulle $H_0$, écartée au profit de l'alternative $H_1$, notée généralement par $W$. Celle-ci est déterminèe par la relation :

$$P(W/H_0) = \alpha.$$

Le complémentaire de $W$ est la région d'acceptation, noté par $\overline{W}$, elle est détermine par :

$$P(\overline{W}/H_1) = 1 - \alpha.$$

## 1.3 Tests paramétriques et non-paramétriques

Lorsque l'on réalise des comparaisons de population on que l'on compare une population à une valeur théorique, il existe deux grands famille de tests : les tests paramétrique et les tests non paramétriques. Donc le but de paramétriqiue de mon-

trer une égalité sur certaines paramétriques, il est pour lequel on fait une hypothèse paramétriquesur la distribution des donnée sous $H_0$(distribution normal, distribution de poisson....), les hypothèses du test concernent alors les paramétre de cette distribution, par ailleurs, les tests non paramétrique est un test ne nécessitant pas d'hypothèse sur la distribution des données. Les donnée sont alors remplacées les donnée statistique ne dépendant pas de moyenne et variance des donnée intiales (statistique d'ordre comme les ranges....), on va aborder maintenant quelques types des test non paramétriques.

### 1.3.1 Tests basés sur la distribution

#### Test de Kolmogorov-Smirnov

Le principe de ce test est de comparer la fonction de répartition empirique $\hat{F}(x)$ à la fonction de répartition théorique $F(x)$ spécifie sous $H_0$. Soit $X_1, ...., X_n$ un $n$ variables aléatoires iid d'une loi absolument continue innconnue définie sur un espace de probabilité $(\Omega, A, P)$ à valeur dans $\mathbb{R}$, ayant la fonction de répartition $F(x)$. La fonction de répartition empirique $\hat{F}(x)$ de l'échantillon $X_1, ...., X_n$ est définie pour $t \in \mathbb{R}$ par :

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i \leq t)} = \begin{cases} 0 & X_{(i)} > t \\ \dfrac{i}{n} & X_{(i)} \leq t \leq X_{(n)} \\ 1 & X_{(n)} \leq t \end{cases}$$

où $X_{(i)}$ sont les statistiques d'ordres associées à l'échantillon (rangées par ordre croissant). En d'autres termes, on estime $F(x) = P(X \leq x)$ au moyenne de proportion $\hat{F}(x)$ d'élément de l'échantillon qui sont inférieurs ou égales à $t$.

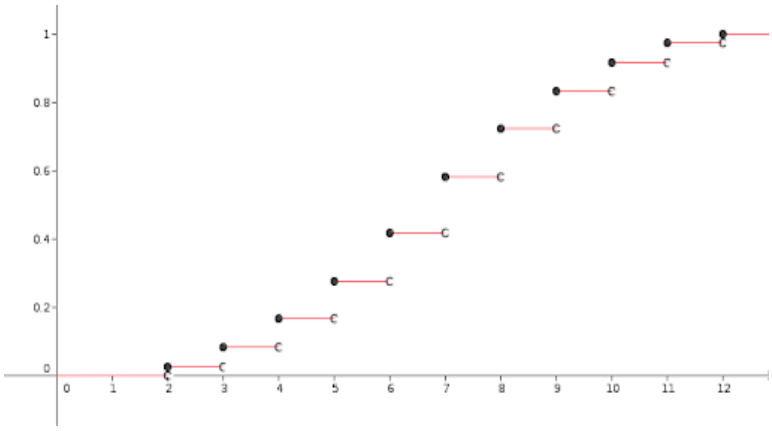


FIG. 1.2 – La fonction de répartition empirique

– On a $\hat{F}(t)$ est une variable à valeurs dans $[0,1]$ .

– On cherche à tester l'hypothèse

$$H_0(P = P_0) \Longleftrightarrow H_0(F = F_0),$$

avec $F_0$ est la fonction de répartition de la loi normal.

– Le théoreme de Glivento-Gantelli donne :

$$\sup \left| \hat{F}(t) - F(t) \right| \underset{n\to\infty}{\to} 0 \quad ps$$

– Il faut donner maintenant un sens à la "distance" entre la fonction de répartition empirique et la fonction de répartition théorique. On mesure l'adéquation de la fonction $\hat{F}(x)$ à la fonction $F(x)$ au moyen d'une distance particulière dite de Kolmogorov-Smirnov, qui est la distance de la norme uniforme entre fonctions de répartitions. Graphiquement, c'est le plus grand écart vertical en valeur absolue entre la valeur empirique et la valeur théorique.

$$KS = D_{KS}(P, P_0) = \sup \mid \hat{F}(t) - F(t) \mid .$$

**Proposition 1.3.1** *Si* $(X_{(1)}, ....., X_{(n)})$ *est la statistique d'ordre de l'échantillon* $X$, *on a :*

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

*On rejette* $H_0$ *si* $D_{KS} > d_{n,\alpha}$, *avec* $d_{n,\alpha}$ *est le quantille théorique*

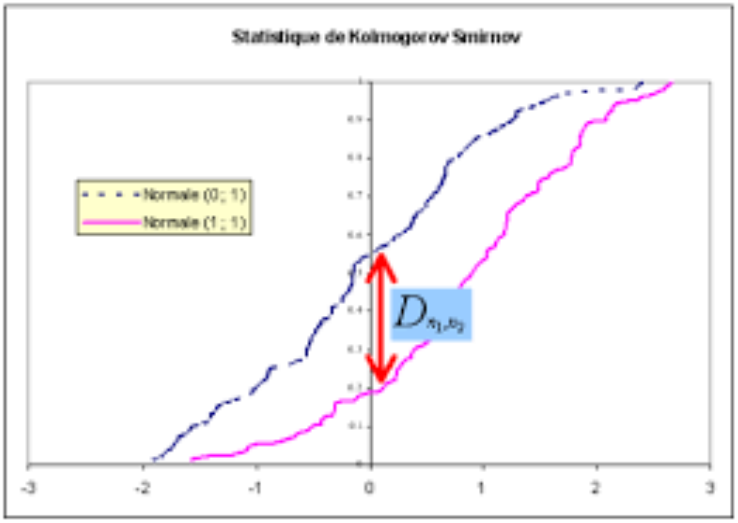


FIG. 1.3 – Détermination de la statistique de Kolmogorov-Smirnov

**Exemple 1.3.1** *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* $H_0(X \backsim Exp(\ \lambda = 1/5)$ *avec un risque* $\alpha = 0.05$.

*Sous R, on utilise la command ks.test du package 'stats' comme suit :*

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

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

*One-sample Kolmogrov-Smirnov test*

*data* $X$ :

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

*alternative hypothesis : two-sided*

$P - value = 0.003$.

*comme la* $p - value$ *est inférieur à la valeurs* $\alpha$*,alors on peut rejetter l'hypothése nulle, c'est à dire accepter* $H_1$.

*Donc, la distribution observée ne suit pas la loi exponentielle de paramètre* $1/5$ *au risque* $5\%$.

**Test de Liliefores** Il s'agit du test de Kolmogorov-Smirnov de la normalité, lorsque la moyenne $\mu$ et l'écart-type $\sigma$ de la distribution normale supposée ne sont pas connus (c'est-à-dire qu'ils sont estimés à partir d'un échantillon de données).

– La statistique de test définie par :

$$\begin{aligned} L_n &= \sqrt{n} KS \\ &= \max_{1 \leq i \leq n} \max \left\{ \left| F_i - \frac{i}{n} \right|, \left| F_i - \frac{i-1}{n} \right| \right\} \end{aligned}$$

où $F_i$ est la fréquence théorique de la loi de répartition normale centrée et réduite associée à la valeur standardisée $Y_{(i)} = \dfrac{X_{(i)} - \overline{X}}{S_x}$ ( 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).

| $D_{crit}$ | $\alpha$ |
|---|---|
| $\frac{0.805}{\sqrt{n}}$ | 0.1 |
| $\frac{0.886}{\sqrt{n}}$ | 0.05 |
| $\frac{1.031}{\sqrt{n}}$ | 0.01 |

TAB. 1.3 – Valeurs critique de test Lilliefors

**Exemple 1.3.2** *Sur un échantillon de taille* 10*, on a observé les valeurs suivantes d'une VD numérique :* $8, 9, 9, 10, 10, 10, 11, 13, 14, 14$.*Est-il légitime de supposer que la distribution de la VD dans la population parente suit une loi normale ?*

*Ce test n'est pas disponible"en standard" avec R, mais il se trouve dans le package "nortest" :*

$>$ *library(nortest)*

$> X(-c(8, 9, 9, 10, 10, 10, 11, 13, 14, 14)$

$>$ *lillie.test*$(X)$

*Lilliefors(Kolmogrov-Smirnov) normality test*

*Data :* $X$

$D = 0.2451$, $p - value = 0.0903$,*donc on Accepte* $H_0(P \sim N(10.8; 4.09))$

**Tests d'Anderson-Darling et de Cramer-von Mises.**

Ces deux débutent avec la même principe que le test KS, à savoir examiner la distance entre la fonction de répartition théorique sous $H_0$ et la fonction empirique construire sur l'échantillon $\hat{F}(x)$. Ils appartiennent à la class des statistiques EDF quadratique (test basé sur la fonction de distribution empirique), si la distribution théorique est $F$, et la fonction empirique est $\hat{F}$, alors les satatistique EDF quadratique mesurent la distance entre $F$ et $\hat{F}$ par :

$$Q = n \int_{-\infty}^{+\infty} (\hat{F}(x) - F(x))^2 \Psi(x) dF(x).$$

où $\Psi(x)$ est fonction de pondération qui va caractériser l'un ou l'autre test.

**Test de Cramer-Von Mises** La fonction de pondération est $\Psi(x) = 1$ et la statistique de test est :

$$W_n^2 = \sum_{i=1}^{n} \left( \frac{2i-1}{2n} - F(X_{(i)}) \right)^2 + \frac{1}{12n}$$

où $X_{(i)}$ est la $i^{ème}$ plus petitw valeur de l'échantillon.

Une grande valeur de la statistique $W_n^2$ est un signe défavorable à $H_0$ : on rejette cette hypothèse si $W^2$ est supérieure à sa valeurs critique.

**Exemple 1.3.3** *On dispose d'un échantillon de n matériels identiques et on note les durées de vie en heures* $x_1, x_2, ..., x_n$. *On a* $n = 5$ *et :*

$$x_1 = 133,\ x_2 = 169,\ x_3 = 8,\ x_4 = 58.$$

*Le paramètre* $x$ *est estimé par* $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 98$, *la fonction de répartition estimé est :*

$$F(x) = 1 - \exp(-\frac{x}{98}),$$

*d'où le tableau :*

| $x_i$ | 8 | 58 | 122 | 133 | 169 |
|---|---|---|---|---|---|
| $F(x_i)$ | 0.079 | 0.447 | 0.711 | 0.743 | 0.821 |

*La statistique de Cramer-Von-Mises vaut :*

$$W_n^2 = \sum_{i=1}^{5} \left( \frac{2i-1}{10} - F(x_i) \right)^2 + \frac{1}{60} = 0.09133$$

*et la quantité* $(1 + \frac{0.16}{n})W_n^2 = 0.0943$. *conduit elle aussi à accepter* $H_0$.

*Sous Rstudio,la commande à utiliser est dans le package goftest :*

$X < -c(8, 58, 122, 133, 169)$.

*cvm.test*$(X)$

*Cramer-vos Mises test of goodness-of-fit*

*Null hypothesis : Uniforme distribution*

*Parameters assumed to be fixed*

*data :X*

$omega2 = 1.667, p - value < 2.2e - 16$

*Comme la* $p-value$ *est inferieur à une risque* $\alpha = 0.05$*.Donc, on accepte l'hypothèse* $H_0$.

**Test d'Anderson-Darling** La fonction de pondération est $\Psi(x) = [F(x)(1 - F(x))]^{-1}$ et la statistique de test s'écrit :

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n} \left[(2i-1)\log F(X_{(i)}) + (2n+1-2i)\log(1 - F(X_{(i)})\right]$$

L'hypothèse de normalité est rejettée lorsque la statistique $A$ prend des valeurs trop élevées :

$$R.C : A > A_{crit}$$

Les valeurs critiques $A_{crit}$ pour différents niveaux de risques sont résumées dans le tableau suivant, ils ont été produits par simulation et ne dépendent pas de l'effectif de l'échantillon :

| $\alpha$ | $A_{crit}$ |
|---|---|
| 0.10 | 0.631 |
| 0.05 | 0.752 |
| 0.01 | 1.035 |

TAB. 1.4 – Valeur critique de test Anderson darling

**Exemple 1.3.4** *Pour une population* $\Omega$*, nous voulons étudier la conformité de la distribution pour une variable aléatoire continue* $X$; *avec la loi normale. Nous disposons cela de* $n_1 = 30$ *observations suivantes :*

$X = (14; 32; 6; 13; 11; 2; 12; 12; 13; 30; 21; 0; 9; 20; 17; 6; 13; 10; 2; 10; 17; 14; 23; 22;$

$13; 21; 18; 16; 27; 20)$.

Est ce que la distribution de échantillon $X$ suit une loi normale ?

Sous R :

$> X < -c(14, 32, 6, 13, 11, 2, 12, 13, 30, 21, 0, 9, 20, 17, 6, 13, 10, 2, 10, 17, 14, 23, 22,$

$13, 21, 18, 16, 27, 20)$.

$>$ ad.test$(X)$

Anderson-Darling normality test

data :$X$

$A = 0.24645, P - value = 0.734.$

On a la $p - value$ est supérieur à un risque $\alpha = 0.05$. Donc on accepte l'hypothèse $H_0$, la distribution de la variable $X$ suite une loi normale.

**Test de Shapiro-Wilk**

Ce test basé sur les $L$-statistiques ( combinaisons linéaires de statistiques d'ordres), qui se base sur une comparaison de la variance empirique avec un estimateur de la variance des $X_i$, qui a de bonnes propriétés sous l'hypothése de normalité :

- Soit $Y_1, ....., Y_n$ i.i.d de loi $N(0,1)$ et $Y_{(1)} \leq ..... \leq Y_{(n)}$ l'échantillon ordonné.
- Soit $\alpha = (E(Y_1), ...., E(Y_n))'$, soit $B$ la matric de covariance de vecteur $(Y_{(1)}, ....., Y_{(n)})$.
- Le test de Shapiro-wilk pour tester l'hypothése de normalité des $X_i$ est basé sur la statistique de test :

$$SW_n = \frac{\hat{\sigma}_n(\alpha' B^{-1}\alpha)^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2(\alpha' B^{-2}\alpha)}$$

avec

$$\hat{\sigma}_n = \frac{\alpha' B^{-1} X_{(i)}}{\alpha' B^{-1}\alpha}.$$

La statistique peut être réecrite sous la forme suivante :

$$SW_n = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

avec :

$$(a_1, ...., a_n) = \frac{\alpha' B^{-1}}{(\alpha' B^{-1} B^{-1}\alpha)^{\frac{1}{2}}}.$$

On rejette $H_0$ si $SW_n \leq c_{n,1-\alpha}$, où les valeurs seuils $c_{n,1-\alpha}$ pour risque $\alpha$ et les effectifs $n$ sont des lues dans la table de Shapiro-Wilk.

**Exemple 1.3.5** *On prend l'exemple précédent*

Sous R :

$X < -c(14, 32, 6, 13, 11, 2, 12, 13, 30, 21, 0, 9, 20, 17, 6, 13, 10, 2, 10, 17, 14, 23,$

$22, 13, 21, 18, 16, 27, 20$).

shapiro.test($X$)

Shapiro-wilk normality test.

data :$X$

$W = 0.9804; p - value = 0.8496$.

On remarque que $p - value$ est supérieure au niveau $\alpha$, ce qui confirme la validité de l'hypothèse $H_0$..Aalors on accepte $H_0$, c'est à dire les données suivent une distribution normale.

**Test de Pearson(khi-2)**

La statistique de Khi-deux ($\chi^2$) est fréquement utilisée en statistique, il est particuliérement adaptée pour les observations qualitatives. Il permet de comparer les fréquences observées sur une ou plusieurs variables aux fréquences théoriques. Elle se base théoriquement sur la loi multinomiale..Soit une épreuve aléatoire ayant $m$ résultats $o_1, ..., o_m$ de probabilités $p_1, ..., p_m$ respectivement avec $\sum_{j=1}^{m} p_j = 1$.Si on le répéte $n$ fois (de maniére indépendantes),on considère $m$ variables aléatoires $X_1, ..., X_m$, désignent le résultat $o_i$ a été obtenu. La fonction de mass conjointe des variables $X_1, ..., X_m$ est donnée par :

$$P(x_1, ..., x_m) = \begin{cases} \dfrac{n!}{x_1!x_2!...x_m} p_1^{x_1} p_2^{x_2} ... p_m^{x_m} & \text{si } x_i \geq 0, \sum_{i=1}^{m} x_i = n \\ 0 & \text{sinon} \end{cases}$$

Alors, on dit que le vecteur $X = (X_1, ..., X_2)'$ suit une loi multinomaile $M(n; p_1, ..., p_m)$.

**Remarque 1.3.1** *La variable aléatoire :*

$$Q = \sum_{i=1}^{m} \frac{(X_i - np_i)^2}{np_i} \sim \chi^2_{(m-1)}, \ \textit{quand } n \to \infty.$$

**Principe général du test du Khi-deux** Soit $O_1, ..., O_m$ des observations d'un vecteur $X = (X_1, ..., X_m)'$ qui suit la loi mutinomiale $M(n; p_1, ..., p_m)$ , on veut tester :

$$\begin{cases} H_0 : P_i = P_i^{(0)} \\ H_1 : P_i \neq P_i^{(0)} \end{cases}$$

où les $p_i^{(0)}, i = 1, ..., m$ sont les valeurs de loi donnée tel que $\sum_{i=1}^{n} p_i^{(0)} = 1$.

– La statistique de test est :

$$\chi_0^2 = \sum \frac{(O_i - np_i^{(0)})^2}{np_i^{(0)}}$$

**Remarque 1.3.2** *1. Lorsque l'hypothèse nulle est vraie, on a :*

$$\chi_0^2 \sim -2\ln(\Lambda'), \text{ quand } n \to \infty$$

*où $\Lambda'$ est la statistique du test de rapport de vraisemblence pour l'hypothèse de test.*

*2. Pour les mêmes conditions, on a :*

$$\chi_0^2 \sim \chi_{(m-1)}^2$$

**Définition 1.3.1 (distance de Khi-deux)** *Il s'agit d'une distance, notée $\chi_0^2$, entre les effectifs observés $(O_i)$ et les effectifs attendus $(np_i^{(0)})$. Lorsque l'hypothèse nulle $H_0$ $(P_i = P_i^{(0)})$ est vraie alors on la rejette si $\chi_0^2$ est grande. On peut effectué ce test par deux façons :*

1. *On rejette $H_0$ si $\chi_0^2 > \chi_{\alpha,m-1}^2$, pour le seuil $\alpha$ donné(où $\chi_{\alpha,m-1}^2$ la valeur théorique lue dans la table de $\chi^2$ à $m-1$ degrés de liberté).*
2. *On calcule le niveau critique observé $p - value = P(\chi_{m-1}^2 \geq \chi_0^2)$, on rejette $H_0$*

*si cette probailité est inférieur à* 0.05.

**Exemple 1.3.6** *Soit A, B et C, trois marques de détergent à lessive. On suppse que la compagnie effectue le sondage auprés de* 500 *répondants et elle obtient les résultats suivants :* 176 *des répondants utilisent la marque A,* 195 *utilisent la marque B,* 81 *utilisent la marque C et* 48 *n'utilisent aucune des trois marques..Soit* $X_1, X_2, X_3$ *et* $X_4$ *le nombre de répondants qui affirment utiliser respectivemnt la marque A, B, C ou autre de détergent. Le vecteur* $X = (X_1, X_2, X_4)'$ *est distribué selon une loi multinomiale de dimension* 4*, de paramètre* $n = 500$ *et* $p_1 = 0,30; p = 0,40; p_3 = 0,20; p_4 = 0,10$. *Peut-on conclure que les parts de marché ont changé, au seuil critique* $\alpha = 5\%$?

*On distingue* $m = 4$ *catégories : A, B, C et autre. On test alors*

$$H_0 : p_1 = 0,30;\ p_2 = 0,40;\ p_3 = 0,20;\ p_4 = 0,10.$$

*Le tableau suivant donne les effectifs observés et les effectifs attendus.*

| | *A* | *B* | *C* | *Autre* | *Total* |
|---|---|---|---|---|---|
| *Effectifs observés* $O_i$ | 176 | 195 | 81 | 48 | 500 |
| *Effectifs attendus* $E_i$ | 150 | 200 | 100 | 50 | 500 |

*En déduit que*

$$\chi_0^2 = \frac{(176-150)^2}{150} + \frac{(195-200)^2}{200} + \frac{(81-100)^2}{100} + \frac{(48-50)^2}{50} = 8,32.$$

*On a le nombre de degré de liberté* $4-1=3$ *et* $\chi^2_{0,05;3} = 7,81$. *Donc on rejette H, et on peut conclure que les parts marché ont effectivement changé.*

### Test de Kuiper

Le test de Kuiper est utilisé pour vérifier si une distribution donnée, ou une famille de distributions est contredite par des preuves provenant d'un échantillon de données. Il porte le nom du mathématicien néerlandais Nicolaas Kuiper.

Le test de Kuiper est étroitement lié aux test de Kolmogrov-Smirnov, plus connue. Les statistiques d'écart $D^+$ et $D^-$ répresentent les tailles absolues des différences les plus positives et les plus négatives entre les deux fonctions de distributions cumulatives comparées. L'astuce du test de Kuiper est d'utiliser la quantité $D^+ + D^-$ comme statistique de test. Ce petit changement rend le test de Kuiper sensible dans les queues que dans la moyenne. Il le rend également constant lors des transformations périodiques de la variable indépendante. Le test d'Anderson-Darling est un autre test qui fournit une sensibilité égale aux queues et à la médiane, mais il ne fournit pas l'invariance cyclique.

**Définition 1.3.2** *La statistique de test ,V, pour le test de Kuiper est définie comme suit..Soit $F$ la fonction de distribution cumultative continue qui doit être l'hypothèse nulle. Désignons l'échantillon de donné qui sont des réalisations indépendantes de variable aléatoire, ayant $F$ comme fonction de distribution ,par $X_i$ avec $i = 1, ..., n$, puis on définit :*

$$\begin{aligned} Z_i &= F(x_i), \\ D^+ &= \max(\frac{i}{n} - Z_i), \\ D^- &= \max(Z_i - \frac{i-1}{n}), \\ V &= D^+ + D^-. \end{aligned}$$

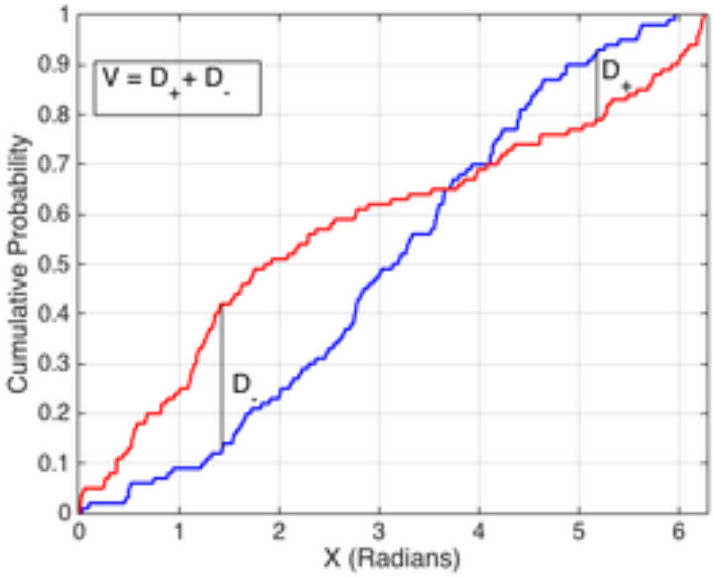


FIG. 1.4 – Visualisation du test Kuiper_2Sample

*Il est autant sensible aux écarts entre les caractéristiques de tendance centrale (ex. la médiane) qu"aux écarts entre les queues de distributions.*

*La région critique du test correspond aux grandes valeurs de $V$. La distribution asymptotique de la statistique est définie par :*

$$P(Kuiper > V) = Q(V \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}})$$

*En pratique, l'intérêt du test de Kuiper par rapport au test de Kolmogorov-Smirnov reste quand même marginal.*

### 1.3.2 Test basé sur les Rangs

#### Test de Wilcoxon

La majorité des tests non-paramétriquse reposent sur les rangs des observations. L'idée est de substituer aux valeurs leurs numéros 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 comparer (étudier l'homogénéité les distributions de $X$ et de $Y$ :

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

On appelle échantillon complet $Z = (Z_1, ...., Z_{n+m}) = (X_1, ...., X_n, Y_1, ...., Y_m)$. On définit $(R_1, ...., R_n)$ comme les rangs de $(X_1, ...., X_n)$ dans l'échantillion complet réeordonné $Z_{(.)}$.comme l'échantillon ne contient pas d'ex-aequo, on a :

$$R_i = \sum_{j=1}^{n+m} \mathbf{1}_{Z_j \leq X_i}$$

– Le statistique de Wilcoxon(1945) est définie par :

$$W_X = \sum_{i=1}^{n} R_i$$

Alors,

$$E(W_X) = \frac{n(n+m+1)}{2},\ Var(W_X) = \frac{nm(n+m+1)}{12}.$$

Sous $H_0(F_X = F_y)$

$$\frac{W_X - E(W_X)}{\sqrt{Var(W_X)}} \sim N(0,1).$$

– La région critique du test au niveau de signification $\alpha$ est $\left|\frac{W_X - E(W_X)}{\sqrt{Var(W_X)}}\right| > z_{1-\frac{\alpha}{2}}$, où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de loi normale centrée réduite.

**Exemple 1.3.7** *La taille des feuille de ronces ont été mesuré pour voir si il y a une différence entre la taille des feuilles qui poussent en plein soleil et qui poussent à*

*l'ombre. Les résultats sont les suivants (largeur des feuills en cm) :*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Soleil* | 6.0 | 4.8 | 5.1 | 5.5 | 4.1 | 5.3 | 4.5 | 5.1 |
| *ombre* | 6.5 | 5.5 | 6.3 | 7.2 | 6.8 | 5.5 | 5.9 | 5.5 |

*On réodonne les* 16 *observations par ordre croissant. Les résultats Soleil sont soulignés :*

*Observations :* $\underline{4.1}\ \underline{4.5}\ \underline{4.8}\ \underline{5.1}\ \underline{5.1}\ \underline{5.3}\ \underline{5.5}\ 5.5\ 5.5\ 5.5\ 5.9\ \underline{6.0}\ 6.3\ 6.5\ 6.8\ 7.2.$

*Rangs :* $1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13\ \ 14\ \ 15\ \ 16;$

*La somme des rangs des individus du Soleil est :*

$$W_X = 1 + 2 + 3 + 4,5 + 4,5 + 6 + 8,5 + 12 = 41.5.$$

*Si* $H_0$ *était vraie :*

$$E(W_X) = \frac{8(8+8+1)}{2} = 68;\ Var(W_X) = \frac{8 \times 8(8+8+1)}{12} = 90.66 = (9.525)^2$$

*Comme* $\frac{41.5-68}{9.525} = -2.7821,$ *et*

$$|-2.7821| \geq z_{0.975} = 1.96(\alpha = 5\%).$$

*alors,on rejette* $H_0$*. La différence entre la taille des feuilles à l'ombre et au soleil est donc significative au risque* $(\alpha = 5\%)$

**Exemple 1.3.8** *Soient* $X$ *et* $Y$ *deux échantillons indépendants à comparer.Sous R, la commande à utiliser est la suivante :*

$> X < -c(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, 0.73, 1.46)$

$> Y < -c(1.15, 0.88, 0.90, 0.74, 1.21)$

$>$ wilcox.test$(X, Y)$

Wilcoxon rank sum test

data : $X$ and $Y$

$W = 5.5, p - value = 0.005907$

Comme $p-value$ est inferieure à la valeur $\alpha = 0,05$,alors on peut rejeter l'hypothése nulle $H_{0.}$

**Remarque 1.3.3 (Traitement des ex-aequos(principe des rangs moyens))**
*Lorsqu'il y a des ex-aequos dans les valeurs (deux ou plusieurs observations présentent la méme valeurs), nous devons définir une stratégie pour effectuer les rangs.*

| valeur | 1.2 | 2.4 | 2.4 | 2.4 | 3.7 | 3.7 |
|---|---|---|---|---|---|---|
| rangs | 1 | 3.0 | 3.0 | 3.0 | 5.5 | 5.5 |

TAB. 1.5 – Traitement des ex aequo-Méthode de rangs moyens

*La méthode des rangs moyens : Il s'agit pour des observations qui prennent la méme valeur. Dans le tableau ,nous effectons* $\frac{2+3+4}{3} = 3.0$ *aux individus correspondant à la valeur* 2.4, *et* $\frac{5+6}{2} = 5.5$ *aux individus correspondant à la valeur* 3.7. *Dans ce cas, aucune modification des tables, lois asyptotiques et son espérance.*

**Test de Mann-Whiteny**

Le test de Mann-Whiteny est un test non paramétrique qui permet de tester si deux échantillons sont issus de populations indépendantes ont la même moyenne. Soit $(x_1, ..., x_n)$ et $(y_1, ..., y_m)$ les deux échantillons de lois respectives $L_X$ et $L_Y$. On test $H_0 : L_X = L_Y$. ou par rapport aux fonctions de distributions $(H_0 : F_X = F_Y)$.

Le principe du test consiste à déterminer le nombre de couples $(X_i, Y_j)$ pour lesquels $X_i \neq Y_j$, et on affecte les rangs de chaque échantillon puis $W$ la somme des rangs de l'échantillon.

– Le test statistique de Mann-Whiteny est défini par :

$$U_{n,m} = \sum_{i=1}^{n}\sum_{j=1}^{m} 1_{(x<y)}(X_i, Y_j).$$

On détermine pour chaque valeur $X_i$ du premier échantillon, le nombre de valeur $Y_j$ de deuxième échantillon telle que $Y_j \geq X_i$. On note $U_1$ la valeur obtenue à partir de premier échantillon et $U_2$ la valeur obtenue à partir du deuxième échantillons :

$$U_1 = R_1 - \frac{n(n+1)}{2} \text{ et } U_2 = R_2 - \frac{m(m+1)}{2}$$

où $R_1$ est la somme des rangs du premier échantillon, et $R_2$ est la somme des rangs de deuxième échantillon. On prend :

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

– On rejette l'hypothèse nulle si $U \in [0, m_\alpha]$, avec $m_\alpha$ d'aprés la table de Mann-Whiteny.

L'espérance et la variance de $U_{n,m}$ s'écrivent :

$$E(U_{n,m}) = \frac{nm}{2},\ Var(U_{n,m}) = \frac{nm(N+1)}{12}.$$

**Remarque 1.3.4** *Pour des échantillon de grande taille ; on a*

$$\frac{U_{n,m} - nm/2}{nm(N+1)/12} \sim N(0,1).$$

**Remarque 1.3.5** *Les statistique des tests de Wilcoxon et Mann-Whiteny sont lieés par la relation suivante :*

$$W_X = U_{n,m} + \frac{n(m+1)}{2}$$

**Exemple 1.3.9** *On prend l'exemple précédent (test de wilcoxon)*

*Les valeurs ordonnées :* 4.1 4.5 4.8 5.1 5.1 5.3 5.5 5.5 5.5 5.5 5.9 6.0 6.3 6.5 6.8 7.

*Les rangs :* 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16.

*Les rangs moyens :* 1 2 3 4.5 4.5 6 8.5 8.5 8.5 8.5 11 12 13 14 15 16.

$$R_1 = 1 + 2 + 3 + 4.5 + 4.5 + 6 + 8.5 + 12 = 41.5.$$
$$R_2 = 8.5 + 8.5 + 8.5 + 11 + 13 + 14 + 15 + 16 = 94.5.$$

*Et*

$$U_1 = R_1 - \frac{n(n+1)}{2} = 41.5 - 36 = 5.5$$
$$U_2 = R_2 - \frac{m(m+1)}{2} = 94.5 - 36 = 58.5$$

*Dans tout les cas on obtient :*

$$U = \min(U_1, U_2) = 5.5$$

*Comme* $U < m_\alpha$,*avec* $m_\alpha = 13$ *d'aprés la table de Mann-Whiteny au risque* $\alpha = 5\%$.*On rejette* $H_0$. *La différence entre la taille des feuilles à l'ombre et au soleil est donc significative au risque* $\alpha = 5\%$.

### Test de la Médiane

Soit $(X_1, ...., X_n)$ et $(Y_1, ...., Y_m)$ deux échantillons indépendants de fonction de répartition $F_0$ et $F_1$ respectivement .On veut tester :

$$\begin{cases} H_0 : F_0 = F_1 \\ H_1 : F_0 \neq F_1 \end{cases}$$

Le test de la médianne consiste à déterminer le nombre de variable $X$ qui sont supérieures à la médiane des observation $(X, Y)$. avec $N = n + m$.

**Définition 1.3.3** *Le test de la médiane est définit par :*

$$M_{n,m} = \frac{1}{n}\sum_{i=1}^{n} 1_{R_i > \frac{N+1}{2}}$$

*où $R_i$ est le rang de la $i^{éme}$ observation de $X$.*

– *Sous $H_0$ et si $N = 2k$ (pair)*:

$$E(M_{n,m}) = \frac{1}{2},\ Var(M_{n,m}) = \frac{n}{4m(N+1)}$$

– *Sous $H_0$ et si $N = 2k + 1$(impair)*:

$$E(M_{n,m}) = \frac{N-1}{2N},\ Var(M_{n,m}) = \frac{n(N+1)}{2mN^2}$$

**Remarque 1.3.6** *Sous $H_0$, la loi de $M_{n,m}$ est hypergeométrique, alors asympotitiqement normale (quand* $\min(n, m) \to \infty$).

***Exemple 1.3.10*** *On souhaite tester l'homogéniété entre deux variables aléatoires*

*$X$ et $Y$, dont on dispose de l'échantillon :*

$$X : 45\ 33\ 38\ 30\ 32\ 47\ 54\ 60\ 82\ 79$$

$$Y : 34\ \ 39\ 29\ 44\ 37\ 62\ 55\ 74\ 101\ 87\ 65$$

*Les couples :* $(X, Y) = 29, \underline{30}, \underline{32}, \underline{33}, 34, 37, \underline{38}, 39, 44, \underline{45}, \underline{47}, \underline{54}, 55, 60, 62, 65, 74, \underline{79}, \underline{82}, 87, 101.$

*Le rang :* $R(X) = 2, 3, 4, 7, 10, 11, 12, 14, 18, 19.$;

*Alors :*

$$M_{n,m} = \frac{1}{10}\sum_{i=1}^{n} 1_{R(X)>11} = \frac{4}{10} = 0,4.$$

*et*

$$\begin{aligned} &\frac{M_{n,m} - E(M_{n,m})}{\sqrt{Var(M_{n,m})}} && = 0.5 < z_{0.975} = 1.96 \\ &\text{avec } E(M_{n,m}) && = \tfrac{N-1}{2N} = \tfrac{20}{42}, Var(M_{n,m}) = \tfrac{n(N+1)}{2mN^2} \quad = 0.023 \end{aligned}$$

*Donc $H_0$ est acceptée (les variables sont homogènes).*

# Chapitre 2

# Test d'indépendance ou d'homogénéité

## 2.1 Test de Kruskal-Wallis

Le test de Kruskal-Wallis répresente les scores du test de Wilcoxon. Il consiste de comparer les distributions de plusieurs populations continues. On dispose de $a$ échantillons indépendants $X_{ij}, j = 1, ..., n_i; i = 1, ..., a$ de tailles respectives $n_1, ..., n_a$, provenant de $a$ populations :

| Echantillon | Observations | | | | | Taille |
|---|---|---|---|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ | ... | $X_{1j}$ | $\cdots$ | $n_1$ |
| 2 | $X_{21}$ | $X_{22}$ | ... | $X_{2j}$ | $\cdots$ | $n_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| $i$ | $X_{i1}$ | $X_{i2}$ | ... | $X_{ij}$ | $\cdots$ | $n_i$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| $a$ | $X_{a1}$ | $X_{a2}$ | $\cdots$ | $X_{aj}$ | $\cdots$ | $n_a$ |

TAB. 2.1 – Tableau des données du test de Kruskal-Wallis

L'hypothèse à tester est :

$$\begin{cases} H_0 : \text{ les distributions des } a \text{ populations sont identiques.} \\ H_1 : \text{ les distributions des } a \text{ populations ne sont pas identiques.} \end{cases}$$

A partir des $a$ échantillons indépendants $X_{ij}, i = 1, ..., a; j = 1, ..., n_i$(supposons $a \geq 3$), pour confronter les hypothèse $H_0$ et $H_1$, on fait les étapes du test de Kruskal-Wallis :

1. Former un seul échantillon de taille $N = \sum_{i=1}^{a} n_i$, et placer les données de l'échantillon ainsi obtenu en ordre croissant.
2. Obtenir les rangs $R_{ij}$ pour l'ensemble des $N$ observations. Lorsque plusieurs observations sont identique s(présence d'ex-aequos), on attribue à celle-ci la moyenne des rangs qu'elles auraient eu si elle avaient été différentes.
3. On calcule la somme et la moyenne des rangs pour chaque échantillon :

$$R_{i.} = \sum_{j=1}^{n} R_{ij}; \overline{R}_{i.} = \frac{R_i}{n_i}; i = 1, ..., a.$$

| Echantillon | Rangs des $N$ observations | | | | | Taille | Somme | Moyenne |
|---|---|---|---|---|---|---|---|---|
| 1 | $R_{11}$ | $R_{12}$ | $\cdots$ | $R_{1j}$ | $\cdots$ | $n_1$ | $R_{1.}$ | $\overline{R}_{1.}$ |
| 2 | $R_{21}$ | $R_{22}$ | $\cdots$ | $R_{2j}$ | $\cdots$ | $n_2$ | $R_{2.}$ | $\overline{R}_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$ | $R_{i1}$ | $R_{i2}$ | $\cdots$ | $R_{ij}$ | $\cdots$ | $n_i$ | $R_{i.}$ | $\overline{R}_{i.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a$ | $R_{a1}$ | $R_{a2}$ | $\cdots$ | $R_{aj}$ | $\cdots$ | $n_a$ | $R_{a.}$ | $\overline{R}_{a.}$ |

TAB. 2.2 – Tableau des calculs relatifs au test de Kruskal-Wallis

La statistique de test est :

$$H^{'} = \frac{12}{N(N+1)} \sum_{i=1}^{a} n_i (R_{i.} - \frac{N+1}{2})^2, \quad (\star) \text{où } N = \sum_{i=1}^{n} n_i.$$

– La région critique : $H^{'} > \chi^2_{\alpha;a-1}$ au seuil critique $\alpha$.

**Remarque 2.1.1** *Lorsque $H_0$ est vraie (la distribution sont identiques), les valeurs critiques de la distribution de la statistique $H^{'}$ peuvent étre déterminées selon des tables. En général, on considère que cette distribution est un Khi-deux avec à $\nu = a-1$ dégré de liberté.*

– si $a = 3$, et $n_i \geq 6, i = 1, 2, 3$;

– ou si $a \geq 4$, et $n_i \geq 5, i = 1, ..., a$.

On peut simplifier l'expression $(\star)$, on retrouve plus couramment la fourmule suivante

dans la littérature :

$$H^{'} = \frac{12}{N(N+1)} \sum_{i=1}^{a} \frac{R_{i.}^2}{n_i} - 3(N+1)$$
$$= \frac{12}{N(N+1)} \sum_{i=1}^{a} n_i \overline{R}_{i.}^2 - 3(N+1).$$

**Remarque 2.1.2** *Lorsque les données comportent des ex-aequo, les rangs sont déterminés comme l'a décrit plus haut. Donc on remplace l'équation* $(\star)$ *par :*

$$H^{'} = \frac{1}{K^2} \left( \sum_{i=1}^{a} \frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4} \right), \ (\star\star)$$

où

$$K^2 = \frac{1}{N-1} \left( \sum_{i=1}^{a} \sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4} \right)$$

**Exemple 2.1.1** *Dans une expérience visant à comparer l'efficacité de quatre méthodess de production d'une huil synthétique, les données d'un indice de pureté suivante ont été obtenus en mesurant six spécimens d"huiles produites avec chacune des quatre méthodes de production* $(A, B, C \text{ et } D)$.

| *Méthode* | *Mesures d'indice de pureté (%)* | | | | | |
|---|---|---|---|---|---|---|
| $A$ | $46,8$ | $47,5$ | $54,2$ | $51,3$ | $48,6$ | $50,3$ |
| $B$ | $51,8$ | $56,7$ | $52,9$ | $57,5$ | $58,1$ | $54,2$ |
| $C$ | $53,8$ | $57,5$ | $58,1$ | $56,7$ | $55,9$ | $57,5$ |
| $D$ | $58,1$ | $66,2$ | $61,0$ | $62,1$ | $57,5$ | $60,4$ |

*La distribution de l'indice de pureté n'est pas la même pour les quatre méthodes de production (à un seuil critique de* $1\%$*).*

*On résume les rangs des observations et la sommes des rangs dans ce tableau.*

| *Méthode* | *Rangs des 24 observations* | | | | | | *Taille* | *Somme* |
|---|---|---|---|---|---|---|---|---|
| $A$ | 1 | 2 | 9, 5 | 5 | 3 | 4 | 6 | 24, 5 |
| $B$ | 6 | 12, 5 | 7 | 15, 5 | 19 | 9, 5 | 6 | 69, 5 |
| $C$ | 8 | 15, 5 | 19 | 12, 5 | 11 | 15, 5 | 6 | 81, 5 |
| $D$ | 19 | 24 | 22 | 23 | 15, 5 | 21 | 6 | 124, 5 |

*Comme on a plusieurs des ex-aequos, on utilise l'équation* $(\star\star)$*. On a* $N = 4 \times 6 = 24$.

$$K^2 = \tfrac{1}{23}(1^2 + 2^2 + 9,5^2 + ... + 15,5^2 \\ +15,5^2 + 21^2 - \tfrac{24\times 25^2}{4}) = 49,652.$$

*Alors*

$$H^{'} = \frac{1}{49,652}(\frac{24,5^2}{6} + \frac{69,5^2}{6} + \frac{81,5^2}{6} + \frac{124,5^2}{6} - \frac{24 \times 25^2}{4}) = 17,028.$$

*Comme* $\chi^2_{0,01;3} = 11,34$ *et* $H^{'} > \chi^2_{0,01;3}$*, on rejette* $H_0$*. La distribution de l'indice de pureté n'est pas la même pour les quatre méthodes de production.*

*Sous R, on peut utiliser la p-value donnée par la fonction"kruskal.test" :*

$> X < -c(46.8, 47, 5, 54.2, 51.3, 48.6, 50.3)$

$> Y < -c(51.8, 56.7, 52.9, 57.5, 58.1, 54.2)$

$> Z < -c(53.1, 57.5, 58.1, 56.7, 55.9, 57.5)$

$> K < -c(58.1, 66.2, 61.0, 62.1, 57.5, 60.4)$

$>$ kruskal.test(list$(X, Y, Z, K)$)

Kruskal- Wallis rank sum test

data : list$(X; Y; Z; K)$

Kruskal-Wallis chi-squared=17.028, $df = 3, p - value = 0.0006973$.

On remarque que la valeur de la statistique $H = 17.028$ (la même valeur qui trouvée manualement), et la $p - value$ est 0.0006973 est inférieur à la valeur de $\alpha = 0.99$ . Alors on rejette l'hypothèse nulle $H_0$ Ceci explique que la distribution de l'indice de pureté n'est pas la même pour les quatre méthodes de production.

## 2.2 Test de Khi-2 :

### 2.2.1 Test de Khi-2 d'homogénéité

Il s'agit ici de demander si deux listes de nombres de même éffectif total peuvent dériver de la même loi de probabilité. L'hypothèse nulle $H_0$ est la suivante : les deux échantillons proviennent de deux variables aléatoires suivent la même loi. La méthode est applicable si chaque unité provenant de l'une des $k$ populations peut étre classée dans une seule des $r$ classes disponibles que l' on note $c_1, ..., c_r$. On note $P_{ij}$ la probabilité de l'unité devennant de la population $i$, avec $i = 1, ..., k$, appartienne à la catégorié $C_j$, avec $j = 1, ..., r$. On cherche à tester :

$$\begin{cases} H_0 : P_{1j} = P_{2j} = ... = P_{kj} = P_j, \text{ pour chaque } j. \\ H_1 : \text{pour au moins un } j, \text{ au moins deux des } P_{ij} \text{ sont différentes.} \end{cases}$$

On suppose qu'on dispose de $m$ échantillons dont un de taille $n_1$ vient de la premiére population, un de taille $n_2$ vient de la deuxiéme, etc. De plus, on suppose que pour les $n_i$ unités vient de la population $i$. On remarque $O_{i1}$ sont de la catégorie 1, $O_{i2}$ de la catégorie 2,...etc. On a :

$$O_{i1} + O_{i2} + ... + O_{ir} = n_i, \ i = 1, ..., k$$

$$\text{et} \quad n_1 + n_2 + ... + n_k = n$$

| | Catégorie | | | | | | |
|---|---|---|---|---|---|---|---|
| population | $C_1$ | $C_2$ | $\cdots$ | $C_j$ | $\cdots$ | $C_r$ | Total |
| 1 | $O_{11}$ | $O_{12}$ | $\cdots$ | $O_{1j}$ | $\cdots$ | $O_{1r}$ | $n_1$ |
| 2 | $O_{21}$ | $O_{22}$ | $\cdots$ | $O_{2j}$ | $\cdots$ | $O_{2r}$ | $n_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$ | $O_{i1}$ | $O_{i2}$ | $\cdots$ | $O_{ij}$ | $\cdots$ | $O_{ir}$ | $n_i$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $k$ | $O_{k1}$ | $O_{k2}$ | $\cdots$ | $O_{kj}$ | $\cdots$ | $O_{kr}$ | $n_k$ |
| Total | $n_{.1}$ | $n_{.2}$ | $\cdots$ | $n_{.j}$ | $\cdots$ | $n_{.r}$ | $n$ |

TAB. 2.3 – Les données du test d'homogénéité

On note :

$$n_i = \sum_{k=1}^{r} O_{ik}, i = 1, ..., k \text{ et } n_{.j} = \sum_{l=1}^{k} O_{lj}, j = 1, ..., r.$$

Lorsque l'hypothèse $H_0$ est vrai, les proportion associées à la classe $C_j, j = 1, ..., r$, et communes à toutes les $k$ populations sont éstimé par :

$$\hat{p}_j = \frac{O_{1j} + O_{2j} + ... + O_{kj}}{n} = \frac{n_{.j}}{n}, j = 1, ..., r.$$

Les effectif attendu donnés par :

$$E_{ij} = n_i \times \hat{p}_j = \frac{n_i \times n_{.j}}{n}, i = 1, ..., k; j = 1, ..., r.$$

– La statistique du test est :

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

– La régle de décision : On rejette l'hypothèse nulle $H_0$ si $\chi_0^2 > \chi^2_{(k-1)(r-1)}$.

**Exemple 2.2.1** *Les données ci-dessous portent sur* 290 *usines qui ont connu des difficultés entre* 1980 *et* 1985 *pour les raisons suivantes : A problème de marché ; B problème financier et C problème d'opérations. Ces usines ont été classées selon leur taille X (nombre d'employé) : P(petite), M(moyenne), G(grand), et la raison principale de leurs difficultés Y. On a obtenu le classement suivant.*

| | $Y$ | | |
|---|---|---|---|
| $X$ | $A$ | $B$ | $C$ |
| $P$ | 34 | 28 | 5 |
| $M$ | 56 | 40 | 33 |
| $G$ | 49 | 18 | 27 |

*Peut-on conclure que les raisons principales des difficultés sont (proportionnellement) les mêmes quelle que soit la taille de l'usine au seuil critique de 5% ?*

*On a* $k = 3$ *et* $r = 3$*, on obtient le tableau des effectifs attendus suivant :*

| | $Y$ | | |
|---|---|---|---|
| $X$ | $A$ | $B$ | $C$ |
| $P$ | $32,11$ | $19,87$ | $15,02$ |
| $M$ | $61,83$ | $38,26$ | $28,91$ |
| $G$ | $45,06$ | $27,88$ | $21,07$ |

*et la statistique du test vaut :*

$$\begin{aligned}\chi_0^2 &= \frac{(34-32,11)^2}{32,11} + \frac{(28-19,87)^2}{19,87} + \frac{(5-15,02)^2}{15,02} + \frac{(56-61,83)^2}{61,83} \\ &+ \frac{(40-38,26)^2}{38,26} + \frac{(49-45,06)^2}{45,06} + \frac{(18-27,88)^2}{27,88} + \frac{(27-21,07)^2}{21,07} \\ &= 16,84.\end{aligned}$$

*On a* $(k-1)(r-1) = 4$ *et* $\chi^2_{0,05;4} = 9,49$,*alors, on rejette* $H_0$. *La raison principale des difficul varie en fonction de la taille de l'usine.*

Sous R :

$> P < -c(34, 28, 5)$

$> M < -c(56, 40, 33)$

$> G < -c(49, 18, 27)$

>tableau< $-matrix(c(P, M, G), 3, 3, byrow = T)$

> tableau

| | [, 1] | [, 2] | [, 3] |
|---|---|---|---|
| [1, ] | 34 | 28 | 5 |
| [2, ] | 56 | 40 | 33 |
| [3, ] | 49 | 18 | 27 |

> chisq.test(tableau)

Pearson's Chi-squared test

data : tableau

$X - squared = 16.841, df = 4, p - value = 0.002075$

Ainsi, la $p-value = 0.0032075$ est inferieur à $\alpha = 0.05$. Donc, on rejtte $H_0$, la raison principal de diffuculté sont les méme quelle que sont la taille de l'usine.

### 2.2.2 Test de Khi-deux d'indépendance

Ce test vérifie l'absence de lien statistique entre deux variables $X$ et $Y$ .Les deux sont dites indépendante lorsqu'il n'existe aucun lien statistique entre elle, autrement dite, la connaissance de $X$ ne permet en aucune maniére de se prononcer sur $Y$. La méthode de Khi-deux permet d'effectuer ce test, particuliérement dans le cas où les variables considérons ne sont pas nécessairement quantitatives. Ainsi, on cherche à tester :

$$\begin{cases} H_0 : X \text{ et } Y \text{ sont indépendante.} \\ H_1 : X \text{ et } Y \text{ sont dépendante.} \end{cases}$$

On considére que chaque variable posséde un nombre fini de modalité, et qu'il y a $c$ types de défauts et $r$ méthodes de production possible. Le test basé sur un tableau de dimension $r \times c$ , dit tableau de contingence. On prend un échantillon de $n$ unité statistique de la population étudiées que l'on classe conjoitement selon les $r$ modalités de la variable $X$ et $c$ modalité de la variable $Y$. Donc, on obtient un tableau de contingence :

| variable $X$ | $y_1$ | $y_2$ | $\cdots$ | $y_j$ | $\cdots$ | $y_c$ | Total |
|---|---|---|---|---|---|---|---|
| $x_1$ | $O_{11}$ | $O_{12}$ | $\cdots$ | $O_{1j}$ | $\cdots$ | $O_{1c}$ | $\sum_{j=1}^{c} O_{1j}$ |
| $x_1$ | $O_{21}$ | $O_{22}$ | $\cdots$ | $O_{2j}$ | $\cdots$ | $O_{2c}$ | $\sum_{j=1}^{c} O_{2j}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_i$ | $O_{i1}$ | $O_{i2}$ | $\cdots$ | $O_{ij}$ | $\cdots$ | $O_{ic}$ | $\sum_{j=1}^{c} O_{ij}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_r$ | $O_{r1}$ | $O_{r2}$ | $\cdots$ | $O_{rj}$ | $\cdots$ | $O_{rc}$ | $\sum_{j=1}^{c} O_{rj}$ |
| Total | $\sum_{i=1}^{r} O_{i1}$ | $\sum_{i=1}^{r} O_{i2}$ | $\cdots$ | $\sum_{i=1}^{r} O_{ij}$ | $\cdots$ | $\sum_{i=1}^{r} O_{ic}$ | $n$ |

Le principe du test du Khi-deux consiste à comparer les effectifs observés $O_{ij}$ aux effectifs $E_{ij}$. Les effectif attendus $E_{ij}$, $i = 1, ..., r; j = 1, ..., c$ sont calculés à partir

des sommes des lignes et des colonnes du tableau de contingence :

$$E_{ij} = \frac{1}{n}\left(\sum_{k=1}^{c} O_{ik}\right) \times \left(\sum_{l=1}^{r} O_{lj}\right), i = 1, ..., r; j = 1, ..., c.$$

La statistique du test est :

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

– La régle de décision : le test consiste à rejeter $H_0$ si $\chi_0^2 > \chi_{\alpha;\nu}^2$, avec $\nu = (r-1) \times (c-1)$.

**Exemple 2.2.2** *Une analyse des pannes d'un certain modèle de composants électroniquess selon la méthode de production $X$ (deux méthode possibles : $M_1$ et $M_2$) et la cause de la panne $Y$ (quatre causes possibles : $A, B, C, D$) a donné les résultats suivants pour 140 composants :*

| | $Y$ | | | |
|---|---|---|---|---|
| $X$ | $A$ | $B$ | $C$ | $D$ |
| $M_1$ | 21 | 48 | 18 | 10 |
| $M_2$ | 6 | 18 | 7 | 12 |

*Peut-on conclure que la cause d'une pannes des composants et la méthode de production sont dépendants à un seul critique $\alpha = 5\%$?*

Tout d'abord, on réalise le tableau de contingence :

| | Y | | | | |
|---|---|---|---|---|---|
| X | A | B | C | D | Total |
| $M_1$ | 21 | 48 | 18 | 10 | 97 |
| $M_2$ | 6 | 18 | 7 | 12 | 43 |
| Total | 27 | 66 | 25 | 22 | 140 |

Ensuit, on construit le tableau des effectifs attendus, à partir des marges du tableau de contingence :

| | Y | | | | |
|---|---|---|---|---|---|
| X | A | B | C | D | Total |
| $M_1$ | 18,71 | 45,73 | 17,32 | 15,24 | 97 |
| $M_2$ | 8,29 | 20,27 | 7,68 | 6,76 | 43 |
| Total | 27 | 66 | 25 | 22 | 140 |

Dans le premier cas, on a $\frac{27\times 97}{140} = 18,71$. La valeur de la statistique du test est

$$\chi_0^2 = \frac{(21-18,71)^2}{18,71} + ... + \frac{(12-6,76)^2}{6,76} = 7,24,$$

où $\nu = (4-1)\times(2-1) = 3$ et $\chi^2_{0,05;3} = 7,81$. On ne rejette pas l'hypothèse nulle, alors la cause d'une panne ne dépend pas de la méthode de production.

Sous R, les commandes associées sont données par $(c = 2; r = 3)$

> A = matrix($c(21, 48, 18, 10, 6, 18, 7, 12)$, nrow =2,byrow = T)

>chisq.test(A)$p.value

> [1] 0.06462645

Comme la $p-value$ est supérieur à la risque $\alpha = 0.05$,on ne rejette pas l'hypothèse $H_0$.

## 2.3 Test de Kolmogrov-Smirnov d'homogénéité

Nous allons construire un test d'homogénéité. Soit $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ de taille $n$ et $m$ respectivement. L'objectif de ce test est de tester si les deux échantillons suivent la méme loi (inconnue)..Si on note par $F_1$ la fonction de répartition de $X_1$ et $F_2$ la fonction de répartition $Y_1$.On chercher à tester $H_0 : F_1 = F_2$ contre $H_1 : F_1 \neq F_2$. Soi les fonctions des répartitions empiriques définies par :

$$\forall t \in \mathbb{R},\ \hat{F}_1 = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_{(i)} \leq t)} \text{ et } \hat{F}_2 = \frac{1}{m}\sum_{j=1}^{m} 1_{(X_{(j)} \leq t)}.$$

On utilise la distance de Kolmogrov-Smirnov suivante :

$$D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup \left| \hat{F}_1(t) - \hat{F}_2(t) \right|$$

– La région critique : on rejette $H_0$ si $D_{n,m} > s_{ksh}$. où $s_{ksh}$ est donnee par latble statistique du test de Kolmogorov-Smirnov (Voia l'annexe)

Nous définissons la quantité

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

La probablilité critique $p$ du test est produite en appliquant la régle suivante

– $0 \leq Z < 0.227$, $p = 1$.

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

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

– $Z \geq 3.1, p = 0$.

**Théor■me 2.3.1** *Avec les hypothèses donnéess ci-dessus on a, sous "$H_0 : F_1 = F_2$"*

$$P(\sqrt{\frac{nm}{n+m}}D_{n,m} \leq \lambda) \to 1 + 2\sum_{k=1}^{+\infty}(-1)^k \exp(-2k^2\lambda^2)$$

**Proposition 2.3.1** *Sous $H_0$ : "$F_1 = F_2$" et si $F_1$ est continue alors la loi de la statistique de $D_{n,m}$ ne dépend pas de F.*

***Preuve.*** *(voir cours G. Turinici page 45).*

***Exemple 2.3.1*** *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 ($n = 9$) et les jeunes ($m = 8$) selon le tablau suivant :*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Vieux* | 19 | 30 | 20 | 10 | 29 | 25 | 21 | 24 | 50 |
| *Jeunes* | 25 | 21 | 17 | 15 | 14 | 14 | 22 | 17 | |

*On va tester l'homogénéité de deux groupes :*

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

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

$> ks : test(vieux; jeune; alternative = "l")$

*Two-sample Kolmogorov-Smirnov test*

*data : vieux and jeune*

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

*alternative hypothesis : the CDF of x lies below that of y*

*Warning message :*

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

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

■

# Conclusion

Nous avons présenté quelques types de tests non paramétriques usuelles, à savoir les tests d'adéquations, les tests de normalité, les tests d'asymétrie, les tests d'homogénéité, etc. Pour chaque test nous avons définit sa statistique ainsi que sa loi de probabilité exacte et sa loi limite. Nous avons aussi définit pour un seuil donner $0 < \alpha < 1$ les régions critiques appropriée. En utilisant le langage R, nous avons présenté les différents packages et les commandes qui corresponds a chaque tests. En outre, nous avons illustré l'application des ces derniers par des exemples concrets.

# Bibliographie


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

[2] Pierre, D., (2015). Cours de Statistiques inférentielles. Licence 2-S4 SI-MASS.

[3] Atmani, F., (2020).Test de Kolmogrov-Smirnov, Mémoire Master en statistique.université de Biskra.

[4] Adjengue,.L., (2014).Méthode statistiques, Concepte, application et exercice.Canada.

[5] Colletaz, G., (2020). Statistique non paramétrique, Econométrie et Statistique Appliquée.

[6] Racotomalala, R., (2011).Test de normalité, Université Lumiére Lyon2.

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

[8] Yahia, D. Cour de MASTER-1, statistique non paramétrique.Université de Mohamed Khider Biskra.

[9] Desgraupes, B. Cours de $L2$, Économie,UNIVERSITÉ PARIS OUEST NANTERRE LA DÉFENSE.

[10] Ruch, J-Jaques, (2013). Staistique : Tests d'hypothèses.

[11] Rahal,.A., (2014). Test de normalité : Simulation en Logiciel R. Mémoire de Master-2 en Statistique Université de Biskra.

[12] Vinatier,S.(2007 − 2008). Comléments de Mathématique, Licence de Biologie,$3^{éme}$ semestre.FACULTÈ DES SCIENCE Et TECHNIQUES DE LIMOGES.

[13] Gabriel ,T.(2019 − 2020).Introduction à la statistique non paramétrique.M1 Math Université Paris Dauphine.

[14] wikipedia, Kuiper's test, htts ://en.wikipedia.org/wiki/kuiper.test.

Annexe A : Logiciel R

## 2.4 Qu'est-ce-que le langage R ?

• R est un logiciel permettant de faire des analyses statistiques et de produire des graphiques. Mais R est également un langage de programmation complet, c'est cet aspect qui fait que R est différent des autres logiciels statistiques. Les informations sur R sont disponibles sur le site :

$$http://www.r-project.org/$$

• 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 diférentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

$P$ : La probabilité.

$H_0$ : L'hypothèse nulle.

$H_1$ : L'hypothése alternative.

$\theta$ : Paramètre.

$W$ : région critique.

$\alpha$ : risque de premier espèce.

$\beta$ : risque de deuxième espèce.

$\overline{W}$ : région d'acceptation.

$\Omega$ : Ensemple de population.

$p - value$ : probabilité critique.

$F(x)$ : La fonction de répartition.

$\hat{F}(x)$ : La fonction de répartition empirique.

$K.S$ :Kolmogrov-Smirnov.

$L_n$ : Statistique de Lilliefore

$\Psi(x)$ : la fonction de pondération.

$W_1^2$ :Statistique de Cramer-Vo- Mises.

$X_{(i)}$ : $i^{ème}$ statistique d'ordre.

$A^2$ : Statistique de Anderson-Darling.

$R.C$ : Région critique.

$SW_n$ :Statistique de test de Shapiro-Wilk.

$i.i.d$ : Indépendant Identiquement Distribués.

$N(0,1)$ : la loi normale standard.

$E(Y)$ : espérance d'une variable aléatoire $X$.

$B$ : la matrice de covariance.

$\chi^2$ : Statistique de Khi-deux.

$V$ : Statistique de Kuiper.

$R(X)$ : le rang de l'échantillon $X$.

$W_X$ :Statistique de Wilcoxon.

$U_{n,m}$ :Statistique de Mann-Whiteny.

$M_{n,m}$ : Statistique de la médiane.

$H^{'}$ : Statistique de Kruskal-Wallis.

## ملخص

في الإحصاء,الاختبار الإحصائي هو إجراء قرار بين فرضيتين. إنها عملية رفض او قبول فرضية. في هذه الرسالة, نقدم لمحة عامة عن الاختبارات اللامعلمية. بشكل عام, نقدم بعض الاختبارات التي تقارن دالة التوزيع التجريبية مقابل دالة التوزيع النظري. على وجه الخصوص, نناقش اختبارات التجانس والاستقلالية بين توزيعين أو أكثر, مع أمثلة باستخدام برنامج تحليل الإحصائي.



## *Abstrac*

*In statistics, a statistical test is a decision procedure two hypotheses. It is a process of rejecting or accepting a hypothesis. In this thesis, we give an overview on the non parametric tests. In general, we present some test that compares the empirical distribution function against the theoretical distribution function. In particular, we discuss the test of homogeneity and independence between two or more distribution, with example using the statistical analysis software "R".*



## *Résumé*

*En statistique, un test statistique est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter où accepter une hypothèse. Dans ce mémoire, on donne un aperçu sur les tests non paramétrique. En général, nous présentons quelque test qui compare la fonction de répartition empirique contre la fonction de répartition théorique. En particulier, nous abordons les tests d'homogénéité et d'indépendance entre deux distribution ou plus ,avec des exemple à l'aide du logiciel d'analyse statistique « R ».*