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

MASTER en Mathématiques

Option : **Statistique**

Par

REDDAS Soumaya

Titre :

Tests de comparaison de deux distributions basée sur la fonction de répartition empirique

Membres du Comité d'Examen :

Dr.	**Sayah Abdallah**	UMKB	Président
Pr.	**Meraghni Djamel**	UMKB	Encadreur
Dr.	**Touba sonia**	UMKB	Examinateur

Juin 2018

Dédicace

Á mon pére : Louardi.

Á ma mére : Aicha

Á mes frères : Walid, Karim, Khaled, et Abdelkader

et ma sœur : Asma

Á Mon fiancé : Fares

Á mes grand mères : Aicha et Zohra.

Á ma belle-famille ;

Á mes collègues et mes Amies plus particulièrement

Mimi, Nassima, Hadda, Rima, Nour, Ibtissam, Nessrin, Bochra, et Houda.

REMERCIEMENTS

Avant tout propos, je tiens à rendre grâce à "Allah" qui m'a guidé sur le bonne voie.

Nous tenons à exprimer notre profonde gratitude à notre directrice de recherche Pr : Meraghni Djamel pour sa disponibilité, sa bienveillance, ses conseils pertinents, son accompagnement et son dévouement.

Nos remerciements vont également aux membres de jury : Sayah Abdallah et Touba Sonia d'avoir accepté d'examiner et d'évaluer notre travail de recherche.

Que Monsieur Benatia Fateh trouve ici l'expression de notre reconnaissance pour leurs conseils ; leurs remarques et leurs orientations.

Table des matières

Table des figures

Liste des tableaux

Introduction

La notion de test d'hypothèse a été développée dans le premier moitié du XXe siècle par un mathématicien anglais "Egon Sharpe Pearson" (1895-1980) et un mathématicien dorigine polonaise "Jerzy Neyman" (1894-1981). L'élaboration des tests d'hypothèse semble avoir ètè rèalisèe en premier lieu au niveau des sciences expérimentales et dans le domaine de la gestion. C'est ainsi que, par exemple, le test de Student qui a ètè dèveloppè par William Sealy Gosset dans le cadre de son activité professionnelle aux brasseries Guinness. Un test d'hypothèse est une procédure permettant de choisir parmi deux hypothèses celles la plus probable au vu des observations effectuées partir d'un échantillon ou un dispositif expérimental, ces deux hypothèses ne jouent pas un rôle symétrique.

Il consiste alors à généraliser les propriétés constatées sur des observations la population d'où ces dernières sont extraites et répondre des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

Dans la théorie des tests d'hypothèse et en pratique, on distingue deux types de tests : un test dit paramétrique si la population mère est de distribution connue et un test dit non paramétrique lorsque la distribution de la population mère est inconnue.

Dans le cadre de ce mémoire on va intéresser à l'utilisation de la fonction de répartition empirique dans la réalisation des tests non paramétriques de comparaison de deux distributions.

Ce mémoire se compose en trois chapitres tels que le premier chapitre a été redigère d'une façon à être lues l'une après l'autre.

Premier chapiter : Fonction de réprtition emirique.

Dans ce chapitre, nous présentons dans un premier temps Fonction de répartition et Statistiques d'ordre. Dans un second temps, nous détaillerons à fonction de répartition empirique (Définition, Propriétés, Graphe et Convergence).

Deuxième chapiter : Tests de comparaison de distribution.

Ce chapitre est dédié aux déffirents tests de comparaison de deux distributions basée sur la fonction de répartition empirique comme le test de Kolmogorov-Smirnov, Test de Cramer-von Mises et le test de Kuiper.

Troisième chapiter : Application sous R.

des applications sur des données simulées ainsi que sur des données réelles l'aide de logiciel R.

On termine notre mémoire par une conclusion.

Chapitre 1

Fonction de répartition empirique

1.1 La fontion de répartition

Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilitè et X une variable aléatoire rèelle sur Ω.

Définition 1.1.1 *[9]on appelle fonction de rèpartion de X la fonction F dènie sur $\mathbb{R}$ par*

$$\forall x \in \mathbb{R}\ F(x) = P(X \leq x). \tag{1.1}$$

Propriété 1.1.1 *La fonction de répartition F vèfirie les properiètès suivantes*

1. $\forall x \in \mathbb{R},\ 0 \leq F(x) \leq 1$.
2. *F est croissante sur $\mathbb{R}$.*
3. *F est continue droite sur $\mathbb{R}$.*
4. $\lim\limits_{x \to -\infty} F(x) = 0$ *et* $\lim\limits_{x \to +\infty} F(x) = 1$.

Preuve. [9]On essais de preuver les propriété précidente :

1. D'apres la définition la fonction de répartition représente une probabilité alros elle est claire que

$$F(x) \in [0, 1].$$

2. Pour : $x \leq y$ on a $]-\infty, x] \subset]-\infty, y]$ et P étant croissante au sens de l'inclusion on a

$$F(x) = P(X \leq x) \leq P(X \leq y) = F(y).$$

Alors $F(x)$ est croissante

3. Soit $A_n = \left]-\infty, x + \frac{1}{n}\right]$, alors $\lim\limits_{n\to+\infty} A_n = \bigcap\limits_n A_n =]-\infty, x]$, de probabilité $F(x)$. On a

$$\begin{aligned}\lim_{y\to x^+} F(y) &= \lim_{n\to+\infty} F(x + \frac{1}{n}) = \lim_{n\to+\infty} P\left(X \leq x + \frac{1}{n}\right) \\ &= \lim_{n\to+\infty} P(X \in A_n) \\ &= P\left(X \in \bigcap_n A_n\right) = F(x).\end{aligned}$$

4. Pour la premiére limite on dit que : pour F est croissant est minorée par 0 elle admet une limite en $-\infty$. De plus soit (a_n) une suite décroissante de réels tendant vers $-\infty$. Alors la suite $A_n =]-\infty, a_n]$ est telle que $A_{n+1} \subset A_n$ et par conséquent $\lim\limits_{x\longrightarrow-\infty} A_n = \bigcap\limits_{n=0}^{+\infty} A_n = \varnothing$. On a

$$\lim_{n\longrightarrow+\infty} F(a_n) = \lim_{n\longrightarrow+\infty} P(X \leq a_n) = \lim_{a\longrightarrow+\infty} P(A_n) = P(\bigcap_{n\in N} A_n) = p(\varnothing) = 0.$$

Alors

$$\lim_{x\longrightarrow-\infty} F(x) = 0.$$

Et pour la même démonstration dans la deuxième limite mais cette fois prend une suite tend vers $+\infty$ et la suite $A_n =]-\infty, a_n]$, où (a_n) une suite croissante de réels tendant vers $+\infty$ montre que :

$$\lim_{x\longrightarrow+\infty} F(x) = 1.$$

■

Remarque 1.1.1 *Voici quelques remarques importantes.*

1. *Si X une variable aléatoire déscret F est une fonction en escalier.*
2. *Si X est continue alors F une fonction continue.*

1.2 Statistiques d'ordre

Soit $(X_1, ..., X_n)$ un échantillon issu d'une loi de probabilité de fonction de répartition F, on a n valeurs observées :

$$(X_1, X_2..., X_n)\,. \tag{1.2}$$

De ce n échantillon. Ordonnons cette suite de valeurs par ordre croissant

$$X_{(n,1)} \leq X_{(n,2)} \leq ... \leq X_{(n,n)}, \tag{1.3}$$

où $(n, 1)$ est le numéro de la plus petite valeur,etc...,(n, n) est le numéro de la plus grande valeur. On note

$$X_{(i)} = X_{(n,i)}. \tag{1.4}$$

La $i - ième$ valeur rangée est la vecteur $\left(X_{(n,1)}, X_{(n,2)}..., X_{(n,n)}\right)$ s'appelle l'échantillon rangé.

Définition 1.2.1 *[6]Pour tout $i = 1, ..., n$ la variable aléatoire $X_{(i)}$ s'appelle la $i - iéme$ statistique d'ordre de l'échantillon*

Exemple 1.2.1 *Les statistiques d'ordre les plus connues sont la plus petite et la plus grande valeur de l'échantillon $X_1 := \min\limits_{1\leq i\leq n} X_i$ et $X_n := \max\limits_{1\leq i\leq n} X_i$. Leurs distributions sont respectivement définies par*

$$F_{\min}(x) = 1 - [1 - F(x)]^n \;\; et \; F_{\max}(x) = [F(x)]^n\,. \tag{1.5}$$

En effet, on a

$$F_{\min}(x) = P\left(X_{(1)} \leq x\right) = 1 - P\left(X_{(1)} > x\right).$$

Puisque $X_{(1)}$ est la plus petite observation alors

$$F_{\min}(x) = 1 - P\left(\underset{1\leq i\leq n}{\cap}[X_i > x]\right) = 1 - \prod_{1\leq i\leq n} P\left([X_i > x]\right) = 1 - [1 - F(x)]^n.$$

Alors que la fonction de répartition de $X_{(n)}$ et de la forme

$$F_{\max}(x) = P\left(X_{(n)} \leq x\right) = P\left(\underset{1\leq i\leq n}{\cap}[X_i \leq x]\right) = \prod_{1\leq i\leq n} P\left([X_i \leq x]\right) = [F(x)]^n.$$

1.3 Fonction de répartition empirique

1.3.1 Définition[11]

En statistique la fonction de répartition empirique est une fonction de répartition qui attribue la probabilité $1/n$ à chacune des n observations d'une variable aléatoire. Soit $(X_1, ..., X_n)$ un échantillon de taille $n \geq 1$ d'une variable aléatoire X de fonction de répartition F. La fonctoin de répartition empirique F_n basée sur l'échantillon $(X_1, ..., X_n)$ est définie par

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)},\ x \in \mathbb{R}, \tag{1.6}$$

où $\mathbf{1}_A$ désigne la fonction indicatrice de l'évenement A. C'est le nombre d'éléments de l'échantillon qui sont inférieurs ou égaux à x. La fonctoin de répartition empirique F_n s'exprime en termes des statistiques d'ordre de la manière suivante

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < x_{(1)}, \\ \dfrac{i}{n} & \text{si} \quad x_{(i)} \leq x < x_{(i+1)},\ i = 1, ..., n-1, \\ 1 & \text{si} \quad x \geq x_{(n)}. \end{cases}$$

1.3.2 Propriétés

Propriété 1.3.1 *La fonction de répartition empirique F_n vérifie les properiétés suivantes :*

1. $\forall x \in \mathbb{R},\ 0 \leq F(x) \leq 1$.
2. F_n *est croissante sur* $\mathbb{R}$.
3. F_n *est continue droite sur* $\mathbb{R}$.
4. $\lim_{x \to -\infty} F_n(x) = 0$ *et* $\lim_{x \to +\infty} F_n(x) = 1$.
5. $F_n(x)$ *est la moyenne empirique des variables aléatoires* $\mathbf{1}_{(X_i \leq x)}$ *qui sont des variables de Bernouilli iid de paramètre* $F(x)$. *Donc* $nF_n(x)$ *est une variable aléatoire binomiale de paramètres* n *et* $F(x)$. *Par conséquent on a*

$$E(F_n(x)) = F(x) \text{ et } V(F_n(x)) = \frac{F(x)(1 - F(x))}{n}. \tag{1.7}$$

6. $F_n(x)$ *est un estimateur sans biais de* $F(x)$ *car* $E(F_n(x)) = F(x)$.

1.3.3 Graphe de F_n

L'application $x \to F_n(x)$ est une fonction en escaliers fonction de répartition de la loi de probabilité uniforme sur l'ensemble $\{x_1, ..., x_n\}$ .Une distribution de probabilité suite une loi uniforme lorsque toutes les valeurs prises par la variable aléatoire sont équibrobable.Si n est le nombre de valeurs différentes prise par la variable aléatoire :

$$P(X = x_i) = \frac{1}{n}, i = 1, ..., n. \tag{1.8}$$

Avec

$$E(X) = \frac{n+1}{2} \text{ et } Var = \frac{n^2 - 1}{12}. \tag{1.9}$$

Et sa graphe est donnée dans la figure [6].

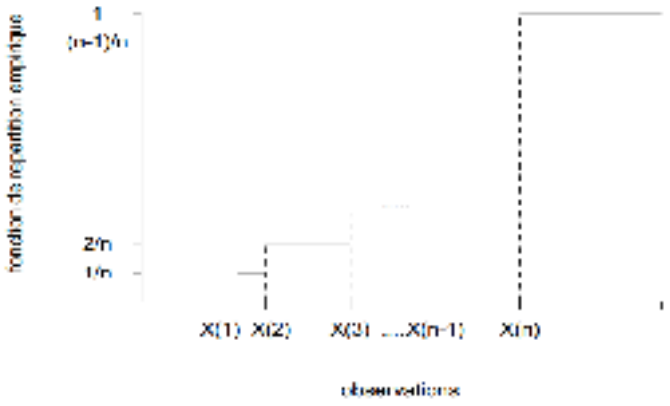


FIG. 1.1 – Fonction de répartition

1.3.4 Convergence de F_n[5]

Soit $x \in \mathbb{R}$ fixé, on a les asympotiques suivantes :

1. Consistance :

$$F_n(x) \xrightarrow{P} F(x) \text{ quand } n \to +\infty. \tag{1.10}$$

2. Consistance forte :

$$\forall x \in \mathbb{R} : F_n(x) \xrightarrow{p.s} F(x) \text{ quand } n \to +\infty. \tag{1.11}$$

3. Théorème de Glivenko-Cantelli :

$$\sup_{x \in R} |F_n(x) - F(x)| \xrightarrow{p.s} 0 \text{ quand } n \to +\infty. \tag{1.12}$$

4. Normalité asympotique :

$$\sqrt{n} \frac{F_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} \to \mathcal{N}(0,1) \text{ quand } n \to +\infty. \tag{1.13}$$

Preuve.

1. On montre que

$$\forall \varepsilon > 0, P\left(\mid F_n\left(x\right) - F(x) \mid \geq \varepsilon\right) \underset{n \to +\infty}{\to} 0.$$

   Pour cela, on applique l'inégalité de Tchebychev

$$\forall \varepsilon > 0, P\left(\mid F_n\left(x\right) - F(x) \mid \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} Var(F_n(x)).$$

   Comme

$$Var(F_n(x)) = \frac{F\left(x\right)\left(1 - F\left(x\right)\right)}{n} \to 0 \text{ quand } n \to +\infty,$$

   alors on obtient le résultat. On note que le même résultat peut être obtenu en appliquant la loi faible des grandes nombres.

2. On applique la loi forte des grandes nombres. On a

$$F_n\left(x\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)} \text{ et } E\left[\mathbf{1}_{(X_i \leq x)}\right] = F(x).$$

3. Pour la démonstration de ce résultat, voir [1].

4. On applique le théorème de Moivre-Laplace qui dit que si (X_n) est une suite de variables aléatoires de loi $\mathcal{B}\left(n, p\right)$, alors

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{L} \mathcal{N}\left(0, 1\right).$$

   On pose $X_n = nF_n(x)$ et $p = F(x)$ et on obtient le résultat.

■

Ceci termine la vérification des propriétés asymptotiques de la fonction de répartition empirique.

1.3.5 Inverse généralisée[12]

Soit F une fonction de répartition. On définit l'inverse généralisée F^{-1} de F par :

$$F^{-1}(x) = \inf\{t \in \mathbb{R}, F(t) \geq x\}, \forall x \in [0,1]. \tag{1.14}$$

Si U une variable aléatoire de loi uniforme sur $[0,1]$ et F une fonction de répartition et son inverse généralisée. La variable aléatoire $X = F^{-1}(U)$ pour fonction de répartition F.

Chapitre 2

Tests de comparaison de distribution

Les tests de comparaison de distributions sont généralement des tests "non paramétriques" lorsque la distribution de la population mére est inconnue. On dispose de deux èchantillons indèpendants $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$ de taille n et m et de lois F et G respectivement, lois supposées toutes deux continues. Notons F_n et G_m les fonctions de répartition empiriques associées à ces échantillons et on se pose la question pour savoir si ces deux populations sont issues de la même loi de probabilité. Pour cela il existe plusieurs tests, nous allons prèsenter seulements les tests basée sur la foction de répartition empirique.

2.1 Test de Kolmogorov-Smirnov[13]

Le test de Kolmogorov-Smirnov consiste à comparer la fonction de répartition empirique du premier échantillon à celle du deuxiéme.

Pour deux échantillons indèpendants $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$ de taille n et m et de lois F et G Le test de Kolmogorov-Smirnov de l'hypothèse H_0 : "$F(x) = G(x)$" contre H_1 : "$F(x) \neq G(x)$" est construit à partir de la statistique :

$$D_{n,m} = \sup_{x \in \mathbb{R}} \mid F_n(x) - G_m(x) \mid . \tag{2.1}$$

Cette statistique représente une distance [14]

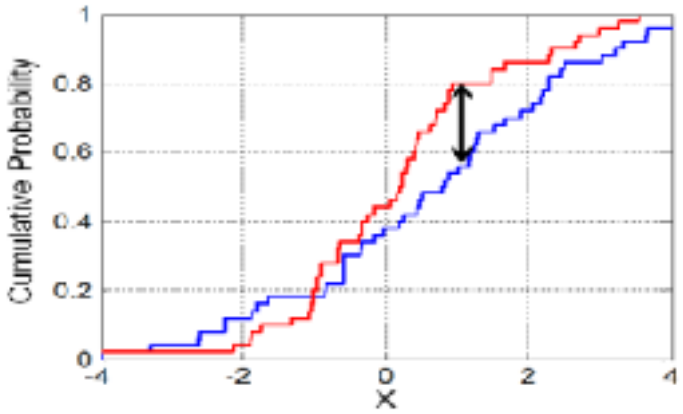


FIG. 2.1 – La représentation graphique de KS

Le fleche noire est exactement la statistique de KS.

La région critique du test pour un seuil α est définie par :

$$\omega = \{D_{n,m} > \lambda\}, \tag{2.2}$$

où λ est telle que :

$$P\left(D_{n,m} > \lambda\right) = \alpha. \tag{2.3}$$

Il consiste à rejeter l'hypothèse H_0 si $D_{n,m} \geq d_{n,m,1-\alpha}$.

2.1.1 Loi de la statistique $D_{n,m}$[4]

L'expression de la loi exacte de la statistique $D_{n,m}$ sous l'hypothèse H_0 est assez simple lorsque $n = m$.

$$P(D_{n,n} \geq d) = 2\sum_{j=1}^{[n/k]} (-1)^{k+1} \frac{(n!)^2}{(n-jk)!\,(n+jk)!}, \tag{2.4}$$

où $d = n/k$, k étant un entier strictement positift $[n/k]$ dèsigne la partie entiére de n/k .

Lorsque $n \neq m$, l'expression de cette loi est beaucoup plus comeplexe et pour des grandes

valeurs de m et n, on peut utiliser la loi asymptotique de $D_{n,m}$. Cette derniére est due à Kolmogorov(1933) et Smirnov(1939). On a :

$$\lim_{n,m\to+\infty} P\left(\sqrt{\frac{nm}{n+m}}D_{n,m} \leq d\right) = 2\sum_{k=1}^{+\infty}(-1)^{k+1}\exp\left(-2k^2d^2\right). \tag{2.5}$$

Remarque 2.1.1 *voici quelque remarque*

1. La statistique de Kolmogorov-Smirnov est aussi définie par :

$$D_{n,m} = \max\left(D^+_{n,m}, D^-_{n,m}\right) \tag{2.6}$$

   Tels que :

$$\begin{aligned} D^+_{n,m} &= \sup_{x\in\mathbb{R}}\left(F_n(x) - G_m(x)\right). \\ D^-_{n,m} &= \sup_{x\in\mathbb{R}}\left(G_n(x) - F_m(x)\right). \end{aligned} \tag{2.7}$$

2. Pour faire un test unilatéral à droit H_0 : "$F(x) = G(x)$" contre H_1 : "$F(x) > G(x)$", sous l'hypothése H_0, on utilise la statistique de test

$$D^+_{n,m} = \sup_{x\in\mathbb{R}}\left(F_n(x) - G_m(x)\right). \tag{2.8}$$

2.2 Test de Cramér-von-Mises

Le test de Cramér-von Mises est aussi basé sur la fonction de répartition empirique. Alors pour tester l'hypothèse H_0 : "$F(x) = G(x)$" contre H_1 : "$F(x) \neq G(x)$"on doit calculer la distance entre deux fonctions et contrairement à (Kolmogorov-Smirnov) qui ce fait par un écart maximal, on utilise la somme des carrés des différences des écarts calculés sur la totalité des observations, la statistique de ce test est donnée par[2] :

$$CvM_{n,m} = \frac{nmS^2_d}{n+m}, \tag{2.9}$$

où S_d^2 est la somme des carrés des différences des écarts définie par :

$$S_d^2 = \frac{1}{n+m} \sum_{i=1}^{n+m} \left(F\left(x_i\right) - G\left(x_i\right)\right)^2 . \tag{2.10}$$

La région critique du test pour un seuil α est définie par :

$$\omega = \left\{CvM_{n,m} > \lambda\right\}, \tag{2.11}$$

où λ est telle que :

$$P\left(CvM_{n,m} > \lambda\right) = \alpha. \tag{2.12}$$

Il consiste à rejeter l'hypothèse H_0 si $CvM_{n,m} \geq cv_{n,m,1-\alpha}$.

2.2.1 Remarque

1. Pour un test bilatéral, on rejette H_0 au niveau de signification 5% (resp. 1%) si CvM est supérieur à 0.461 (resp. 0.743).
2. En raison de la prise des carrés des écarts, ce test est obligatoirement bilatéral.
3. Le test de Cramer-von-Mises a les mêmes applications que le test de Kolmogorov.La diffèrence entre ces deux tests rèside dans le fait que le test de Cramer von Mises prend en compte la somme des carrès des ècarts obtenus alors que le test de Kolmogorov-Smirnov ne s'appuie que sur la plus grande des diffèrences. Ce dernier est donc plus sensible à l'existence des points aberrants.
4. Le test de Cramér-von Mises est souvent plus puissant que le test de Kolmogorov-Smirnov et il est plus facile à utiliser grace à la bonne approximation qui évite le recours à des tables.

2.3 Test de Kuiper[2]

Une variante assez célèbre du test de Kolmogorov-Smirnov est connue sous le nom de test de Kuiper (Nicolaas Hendrik Kuiper, 1960). Elle consiste à prendre en compte la valeur maximale positive et celle négative de la différence entre les deux courbes de distribution à comparer. Test de Kuiper de l'hypothése H_0 : "$F(x) = G(x)$" contre H_1 : "$F(x) \neq G(x)$". Il est basé sur la statistique de test suivante, dite statistique de Kuiper :

$$K_{n,m} = D^+_{n,m} + D^-_{n,m}. \tag{2.13}$$

Avec :

$$\begin{aligned} D^+_{n,m} &= \sup_{x\in\mathbb{R}} \left(F_n(x) - G_m(x)\right). \\ D^-_{n,m} &= \sup_{x\in\mathbb{R}} \left(G_m(x) - F_n(x)\right). \end{aligned} \tag{2.14}$$

Cette modification permet au test de Kuiper un gain considérable de sensibilité au niveau des queues des distributions et des médianes.

Cette statistique est définie graphiquement comme suite [15] :

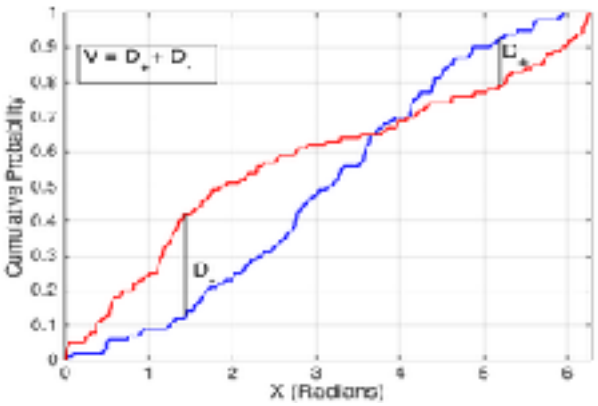


FIG. 2.2 – La réprésentation graphique de kuiper

Propriété 2.3.1 *voici les propriétés suivants :*

1. La loi de la statistique $K_{n,m}$ est tabulée, sous l'hypothèse H_0, ce qui permet de réaliser des tests d'égalité à distance finie.
2. Sous l'hypothèse H_0, le test asymptotique tient compte de la propriété asymptotique suivante :[10]

$$\lim_{\min(n,m)\to+\infty} P(\frac{\sqrt{nm}}{\sqrt{n+m}} D^+_{n,m} > x) = \exp(-x^2) 1_{\mathbb{R}^+}. \tag{2.15}$$

Remarque 2.3.1 *La loi de la statistique $K_{n,m}$ est tabulée et peut être retrouvée en table 54 du volume 2 de l'ouvrage de E.S.Pearson et H. O. Hartley : Biometrika Tables for Statisticians (1972). L'hypothèse reste la même que pour la version de Kolmogorov-Smirnov.[7]*

Chapitre 3

Application sous R

3.1 Sur la fonction de répartition empirique

Exemple 3.1.1 *On considère les données d'un essai visant à déterminer la solidité d'une corde d'escalade. Un morceau de 1 m de la corde est mis sous tension jusqu'à cassure. On se demande si la corde peut casser à n'importe quel endroit. On obtient ainsi les résultats suivants :[8]*

les données d'un essai visant	0.1	0.4	0.4	0.6	0.7	0.7	0.8	0.9	0.9	0.9

TAB. 3.1 – Exemple sur la fonction de répartition empirique

Données réelles(Exemple)

On utilise la fonction de répartition empirique sous logicial R.

$x = c(0.1, 0.4, 0.4, 0.6, 0.7, 0.7, 0.8, 0.9, 0.9, 0.9)$

F =ecdf(x)

$> F(x)$

0.1, 0.3, 0.3, 0.4, 0.6, 0.6, 0.7, 1.0, 1.0, 1.0

library(EnvStats)

ecdfPlot(F(x),type="s",xlab="",ylab="",main="",xlim=c(0,1))

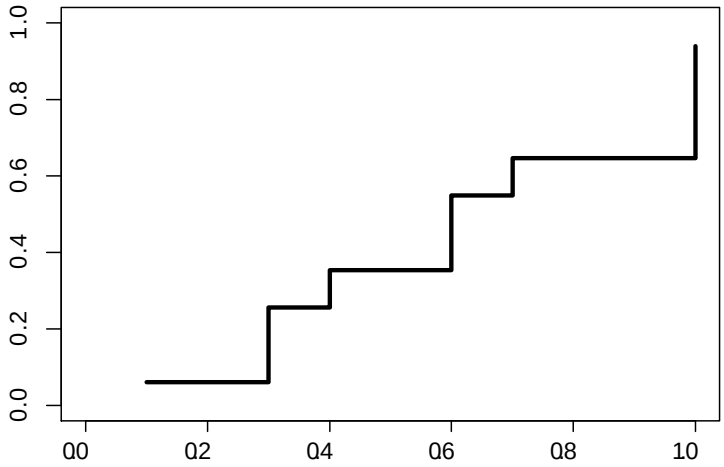


FIG. 3.1 – La fonction de répartition empirique du données

3.2 Tests de comparaison de distribution

3.2.1 Test de kolmogorov-Smirnov

Exemple 3.2.1 *Un psychologue fait passer un test de rapidité à des enfants normaux* (X) *et d'autres considérés comme mentalement retardés* (Y)*. Les temps qu'ils mettent pour accomplir une série de tâches sont les suivants :[8]*

Enfants normaux (X)	183	202	197	204	218	227	233	
Enfants retardées (Y)	202	220	228	239	242	243	261	270

TAB. 3.2 – Exemple sur le test de Kolmogorov-Smirnov

La répésentation graphique de ces données est :

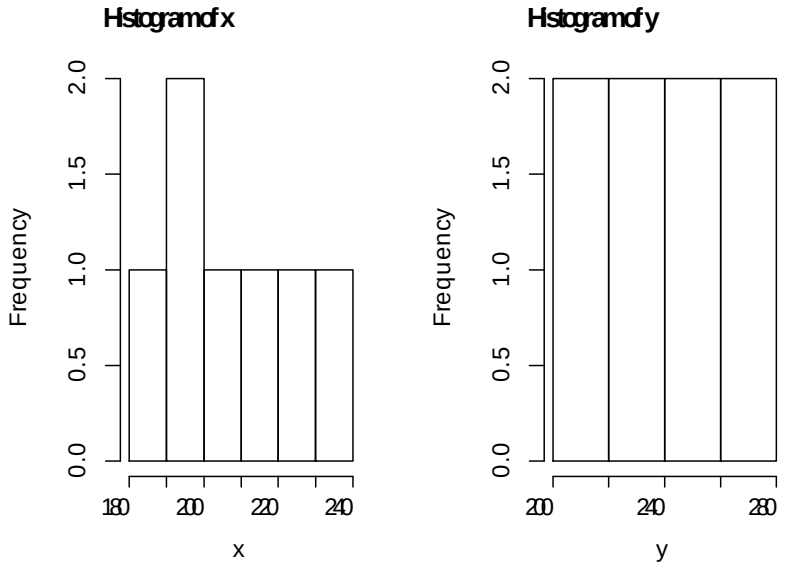


FIG. 3.2 – la répresentation graphique de les donnés

On pose les hypothése de test

$$H_0 : F_n = G_m \text{ contre } H_1 : F_n \neq G_m.$$

La mise en oeuvre du test conduit à :

1. Classer par ordre croissant les valeurs de X et de Y.
2. Calcule des valeurs numériques des fonctions de répartition F_n et G_m de X et Y respectivement.
3. En déduire, point par point, l'expression des différences $\mid F_n(x) - G_m(x) \mid$.
4. Construire cette statistique $\sup_{x \in \mathbb{R}} \mid F_n(x) - G_m(x) \mid$.
5. Dans le tableau ci-dessous N_X et N_Y désignent, pour chaque valeur de x ,le nombre de valeurs de X (resp.de Y) inférieures ou égales à x , les rapports $\frac{N_X}{7}$ et $\frac{N_Y}{8}$ étant en conséquence les valeurs de F_n et de G_m au point x.

les valeurs de x	N_X	N_Y	$F_n(x) = \frac{N_X}{7}$	$G_m(x) = \frac{N_Y}{8}$	$\mid F_n(x) - G_m(x) \mid$
183	1	0	1/7	0	1/7
191	2	0	2/7	0	2/7
197	3	0	3/7	0	3/7
202	3	1	3/7	1/8	17/56
204	4	1	4/7	1/8	31/56
218	5	1	5/7	1/8	33/56
220	5	2	5/7	2/8	26/56
227	6	2	6/7	2/8	34/56
228	6	3	6/7	3/8	27/56
233	7	3	1	3/8	5/8
239	7	4	1	4/8	1/2
242	7	5	1	5/8	3/8
243	7	6	1	6/8	2/8
261	7	7	1	7/8	1/8
270	7	8	1	1	0

TAB. 3.3 – Les rèsultats obtenus

La distance $D_{n,m} = 5/8 = 0.625$ et la table de Kolmogorov-Smirnov fournit pour $n = 7$ et $m = 8$ et $\alpha = 0.05$, la valeur $\lambda = 0.71$.

La loi du supremum des écarts en valeur absolue est tabulée dans la table de Smirnov. On rejette H_0 si $D_{n,m} > 0.71$ donc ici, on ne peut pas rejeter l'hypothèse selon laquelle les distributions sont diérentes pour les deux groupes d'enfants.

Données simulées

On utilise la fonction du test de Kolmogorov-Smirnov sous logiciel R définie précédemment pour comparer un échantillon de taille 500 d'une loi uniforme avec un échantillon de taille 500 d'une loi normal de la manière suivante :

>library(stats) ; X = rexp(500) ; Y = rnorm(500) ; ks.test(X, Y)

Commentaire

On constate que la valeur $p = 0$ est inférieure au seuil $\alpha = 0.05$, ce qui nous permet de conclure que les deux échantillons ne proviennent pas de la même population.et $D = 0,5$

Données réelles(Exemple)

>Enfants normaux= $c\,(183, 191, 197, 204, 218, 227, 233)$;

>Enfants retardées= $c\,(202, 220, 228, 239, 242, 243, 261, 270)$;

>Ks.test(Enfants normaux , Enfants retardées)

Commentaire

On constate que la valeur $p = 0.056$ est superieur au seuil $\alpha = 0.05$, ceci permet de confirmer le résultat obtenu dans l'exemple et $D = 0.625$

On remarque qu'on trouve le même resultat on utilisant la réprésentation graphique.

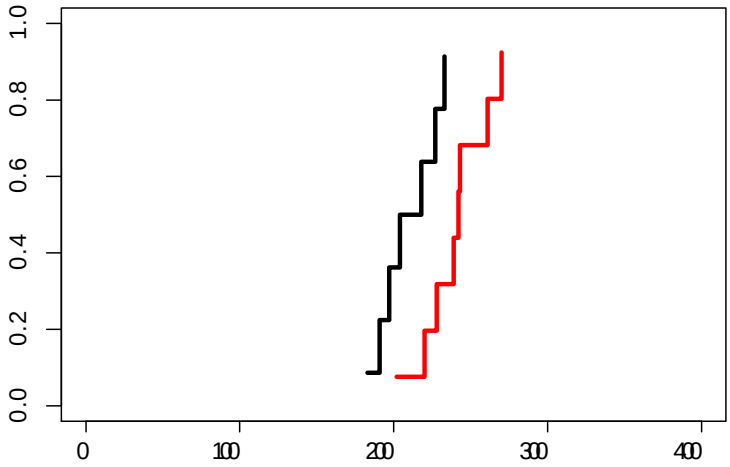


FIG. 3.3 – les fonctions de répartitions empiriques de X et Y

3.2.2 Test de Cramér-von Mises

Exemple 3.2.2 *Par l'observation de la taille du domaine vital (en km^2) pour neuf ours femelles (X) et six ours mâles (Y), on teste l'hypothèse H_0 : " il n'y a pas de différence signicative dans les tailles du domaine vital pour les ours mâles et femelles", contre l'hypothése H_1 : "la taille du domaine vital pour les ours mâles différe signicativement de celle du domaine vital pour les ours femelles". Les domaines vitaux (en km^2) ainsi observès pour les ours considèrès sont les suivants :[1]*

Ours femelles (X)	37	72	60	49	18	50	102	49	20
Ours mâles (Y)	94	504	173	560	274	168			

TAB. 3.4 – Table Caption

La réprésentation graphique des données

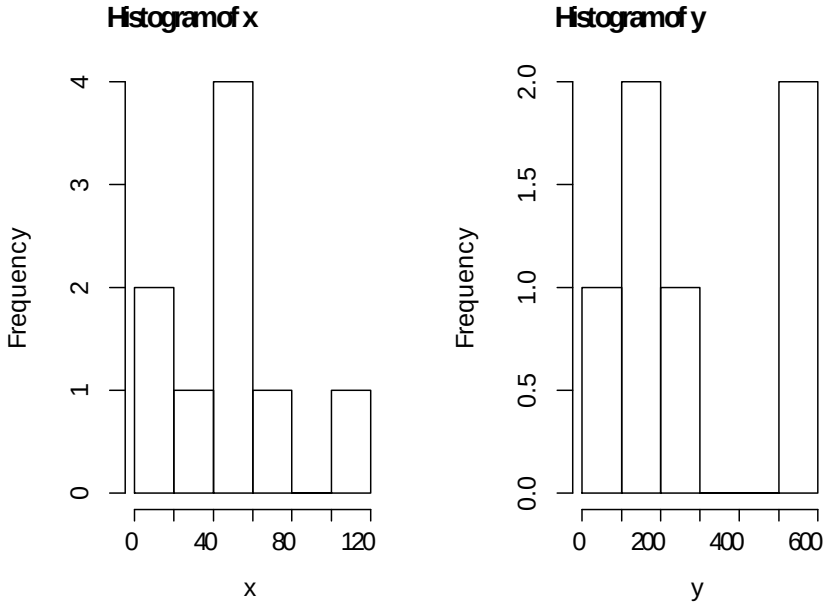


FIG. 3.4 – la répresentation graphique de les donnés

Ce test est construit par les etapes suivants :

1. Classer par ordre croissant les valeurs de X et de Y.

2. Calcule des valeurs numériques des fonctions de répartition F_n et G_m de X et Y respectivement.
3. En dèduire, point par point, l'expression des diffèrences carres$(F_n(x) - G_m(x))^2$.
4. Calculer la statistique $\frac{nm}{(n+m)^2} \sum_{i=1}^{n+m} (F(x_i) - G(x_i))^2$.

Dans le tableau ci-dessous N_X et N_Y dèsignent, pour chaque valeur x, le nombre de valeurs de X (resp. de Y) infèrieures ou ègales x, les rapports $N_X \backslash 9$ et $N_Y \backslash 6$ ètant en consèquence les valeurs de F_n et de G_m au point x :

valeurs x	N_X	N_Y	$F_n(x) = \frac{N_X}{9}$	$G_m(x) = \frac{N_Y}{6}$	$(F_n(x) - G_m(x))^2$
18	1	0	0.111	0	0.012321
20	2	0	0.222	0	0.049284
37	3	0	0.333	0	0.110889
49	5	0	0.555	0	0.308025
50	6	0	0.666	0	0.443556
60	7	0	0.777	0	0.603729
72	8	0	0.888	0	0.788544
94	8	1	0.888	0.166	0.521284
102	9	1	1.00	0.166	0.695556
168	9	2	1.00	0.333	0.444889
173	9	3	1.00	0.500	0.25
274	9	4	1.00	0.666	0.1156
504	9	5	1.00	0.833	0.027889
560	9	6	1.00	1.00	0

TAB. 3.5 – Les rèsultats obtenus

La distance $CvM_{n,m} = 1,042$ et pour $n = 6; m = 9$ et $\alpha = 0,05$;et puisque on a $1,042 > 0,461$ alors ceci conduit donc a rejeter l'hypothése H_0.

Données simulées

Le test de Cramér-von Mises est mis en œuvre dans R par la fonction CvM.test() sous le package "RVAideMemoire". Pour illustrer l'utilisation de cette fonction, on s'intèresse à l'exemple suivant :

On prend un échantillon de taille 30 d'une loi de Poisson de pârametre $\lambda = 2$ avec un autre échantillon de la meme taille d'une loi Poisson aussi mais de paramâtre $\lambda = 3$.
Pour tester avec Cramér-von Mises :
>set.seed(1109) ;x = rpois(30,2) ; y <- rpois(30,3) ; CvM.test(x,y)

Commentaire

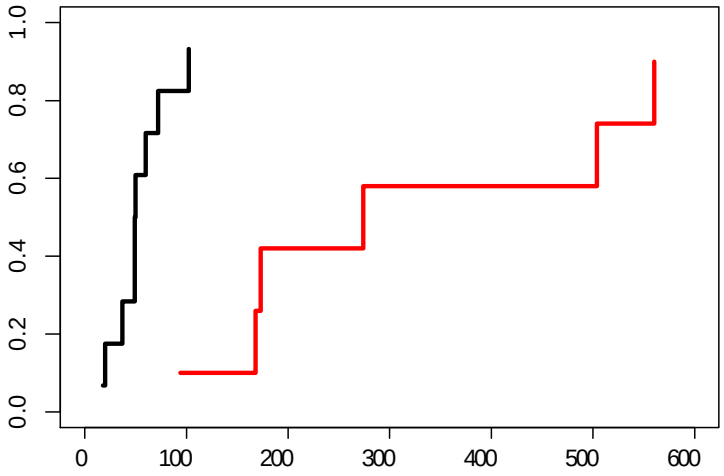
Les résultats du test de Cramer-von Mises confirme que les deux échantillons ne suivent pas la même loi, car la valeur $p = 0,003$ est inferieur au seuil $\alpha = 0,05$.

Données réelles(Exemple)

>Ours femelles $= c\,(37, 72, 60, 49, 18, 50, 102, 49, 20)$;
>Ours males= $c\,(94, 504, 173, 560, 274, 168)$;
>CvM.test(Ours femelles , Ours mles)

Commentaire

La valeur$p = 0.056$ est superieur au seuil $\alpha = 0.05$, ceci permet de confirmer le résultat obtenu pour cet exemple.
Ce resultat est confirmé graphiquement par les réprésentations des fonctions de répartition empirique de X et Y



les fonctions de répartitions empiriques de X et Y

Remarque 3.2.1 *On essaie de tester sous R cette exemple avec le test de Kolmogorov-Smirnov alors la valeur* $p = 0.007$ *est inferieur à un seuil* $\alpha = 0.05$*, ceci permet de confirmer le resultat*

3.2.3 Test de kuiper

Exemple 3.2.3 *D'aprés l'exemple précédente*

Ce test est construit par les même etapes que le test de cramer-von Mises seuf la quatréme étape tels que on calcule la statistique $(F_n(x) - G_m(x)) + (G_m(x) - F_n(x))$.
Dans le tableau ci-dessous N_X et N_Y dèsignent, pour chaque valeur x, le nombre de valeurs de X (resp. de Y) infèrieures ou ègales x, les rapports $N_X \backslash 9$ et$N_Y \backslash 6$ ètant en consèquence les valeurs de F_n et de G_m au point x.

valeurs x	N_X	N_Y	$F_n(x) = \frac{N_X}{9}$	$G_m(x) = \frac{N_Y}{6}$	$F_n(x) - G_m(x)$	$G_m(x) - F_n(x)$
18	1	0	0.111	0	0.111	-0.111
20	2	0	0.222	0	0.222	-0.222
37	3	0	0.333	0	0.333	-0.333
49	5	0	0.555	0	0.555	-0.555
50	6	0	0.666	0	0.666	-0.666
60	7	0	0.777	0	0.777	-0.777
72	8	0	0.888	0	0.888	-0.888
94	8	1	0.888	0.166	0.722	-0.722
102	9	1	1.00	0.166	0.834	-0.834
168	9	2	1.00	0.333	0.667	-0.667
173	9	3	1.00	0.500	0.5	-0.5
274	9	4	1.00	0.666	0.334	-0.334
504	9	5	1.00	0.833	0.167	-0.167
560	9	6	1.00	1.00	0	0

TAB. 3.6 – Les rèsultats obtenus

la statistique $K_{n,m} = 0.888$ et pour $n = 6, m = 9$ et $\alpha = 0,05$;et puisque on a $K_{n,m} < 1,24$ alors ceci conduit donc a rejeter l'hypothése H_0.

Conclusion

Enfin, nous pouvons dire que les tests de comparaison de distribution ont une grande importance dans tous les domaines scientifiques étaient hydrologiques, biologiques ou autres, car ils nous permettent d'obtenir des informations concernant une population inconnue sur la base d'un ensemble d'observations statistiques provenant de cette population.à cet effet, l'objectif principal de notre mémoire est de passer en revue les différents tests de comparaison de distribution basée sur la fonction de répartition empirique.

Pour ce faire, on va comoncer de parler sur la fonction de répartition empirique, tels que la fonction de répartition, Statistiques d'ordre, la fonction de répartition empirique, la définition, leur Propriétés, leur graphe, et leur convergence.

Et une grande partie de ce travail est consacré aux principaux tests de comparaison de distribution basée sur la fonction de répartition empirique telle que le test de Kolmogorov-Smirnov, de Cramer-Von mise et celui de Kuiper.

L'application de ces tests sur des données réelles et simulées, nous permettrons de conclure que les tests de comparaison de distribution sont très variés et l'utilisation de l'un ou l'autre demande beaucoup d'attention et de vérification des conditions d'application.

Nous vous rappelons que les tests que nous avons abordés dans ce mémoire ne sont pas les seuls utilisés dans l'étude de comparaison de distribution mais il y a de nombreux autres tests qui répondent à des problématiques identiques mais s'appuient avec des mécanismes différents.

Il nous semble intéressant et plus que nécessaires de vérifier que plusieurs échantillons

sont issus d'une même population. Ce test est appelé L'analyse de la variance, il s'applique lorsque l'on mesure une ou plusieurs variables explicatives catégorielles qui ont de l'influence sur la distribution d'une variable continue à expliquer. statistique

Bibliographie

[1] Baudart, G.Chatelain, P. (28 Avril 2010), Test de kolmogorov-Smirnov,Thibault Rieutord.

[1] Boulay, J-P. (2010). Statistique mathèmatique, Applications commentèes. Ellipses, Paris.

[2] Colletaz, G. (2017). Statistique non paramétrique. Lien : www.univ-orleans.fr/deg/masters/ESA/GC/sources/CoursNP.pdf.

[3] Gaudoin, O. (2009). Principes et Méthodes Statistiques (Notes de cours). Ensimag - 2éme année. Lien : *https ://www-ljk.imag.fr/membres/Olivier.Gaudoin/PMS.pdf.*

[4] Gouaned, Y. (2015). Mémoire deTests d'Ajustement et de Comparaison.

[5] Guérin, H. Malrieu, F. (2007) Test de kolmogorov-Smirnov_ Convergence des quantiles, Université de Renne.

[6] Lejeune, M. (2010). Statistique : la théorie et ses applications. Springer, Paris.

[7] Lemakistatheux.wordpress.com/2013/05/09/le-test-de-kolmogorov-smirnov/

[8] Monbet, V. (2009).Tests statistiques "Notes de cours L2 S1 ".

[9] Necir, A. (2016). Cours de troisième année. Université Mohamed Khider de Biskra.

[10] http ://jeanalain.monfort.free.fr/Dicostat2005/T/Test_de_KRUSKAL_WALLIS.pdf.

[11] Saporta, G. (2006). Probabilité, analyse de données et statistique. Technip, Paris.

[12] Thas, O. (2010). Comparing distributions. New York : Springer.

[13] https ://WikiStat.com/.../Tests-non-paramétriques/.

[14] https ://Wikipedia.com/.../test-de-Kolmogorov-Simirnov/

[15] https ://Wikipedia.com/.../test-de-Kuiper/.

Annexe : Abréviations et Notations

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

α	Risque de premier espèce.
$CvM_{n,m}$	Statistique de Cramér-von Mises.
$cvm_{n,m,1-\alpha}$	Valeur critique de Cramér-von Mises.
CvM.test(x,y)	la fonction de test de Cramér-von Mises sous R.
$D_{n,m}$	Statistique de Kolmogorov-Smirnov.
$d_{n,m,1-\alpha}$	Valeur critique de Kolmogorov-Smirnov.
$E(X)$	Espérance mathématique de X.
F	Fonction de répartition.
F_n	Fonction de répartition empirique.
F^{-1}	Fonction des quantilles.
iid	Indépendantes identiquement distribuées.
$K_{n,m}$	Statistique de Kuiper.

K_a	la valeur asymptotique de kuiper
ks.test(x,y)	la fonction de test de Kolmogorov-Smirnov sous R.
v.test(x,y)	la fonction de test de kuiper sous R.
v.a	Variable aléatoire.
$Var\,(X)$	Variance mathématique de X.
$(X_1, X_2..., X_n)$	Échantillon de taille n des v.a's.
$(X_{(n,1)}, X_{(n,2)}, ..., X_{(n,n)})$	Statistiques d'ordre associe à $(X_1, X_2..., X_n)$.
$\mathbf{1}_A$	Fonction indicatrice de l'ensemble A.
$\xrightarrow{P}$	Convergence en probabilité.
$\xrightarrow{L}$	Convergence en loi.
$\xrightarrow{p.s}$	Convergence presque sûrement.
:=	Égalité par définition.