République Algériennne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

*Université Mohamed Khider, Biskra*

Faculté des Sciences Exactes

Département de Mathématiques


Mémoire présenté pour obtenir le diplôme de

**Master en "Mathématiques Appliquées"**

**Option : Statistique**

**Par : Bentrcia Hadil**

**Titre :**

# Ajustement des modèles de probabilités et applications

Devant le Jury :

| | | | | |
|---|---|---|---|---|
| Mr. | NECIR Abdelhakim | Pr. | U. Biskra | Encadrant |
| Mr. | MERAGHNI Djamel | Pr. | U. Biskra | Président |
| Melle. | SOLTANE Louiza | Dr. | U. Biskra | Examinateur |

**Soutenu Publiquement le 03/06/2025**

# *Dédicace*

**Je dédie ce modeste travail** *:*

*À mes parents, qui m'ont bien élevée ont semé l'espoir en moi et a appris à être.*

*Mon papa "**Bentrcia Tayeb**" et ma maman " **Bentrcia Hayet**".*

*À ma sœur "**Ibtihal**".*

*À mes frères "**Seyf-eddine**" ,"**Tadj-eddine**" et "**Abd arahim**".*

*À mes amies les plus chères "**Abir**" et "**Omaima**".*

*À mes proches qui m' ont encouragée avec un cœur sincère "**Samira**" ,"**Lwiza**" et "**Salima**".*

*À ma grand mère "**Aicha**".*

*À ma promotion* **2024/2025**

*Et à mes respectables Professeurs qui m' ont apporté tout le soutien et l'orientation nécessaire.*

# Remerciements

*Avant tout, je remercie Dieu, Le Tout-Puissant, de m'avoir accordé le courage, le moral, la santé et la patience tout au long de mes années d'études, qui m'ont* permis d'achever mon parcours et de réaliser ce modeste travail.

*J'exprime toute ma gratitude à mon encadreur,* ***Professeur Abdelhakim Necir****, pour sa disponibilité, son soutien et ses remarques précieuses qui m'ont permis de présenter ce travail sous sa meilleure forme.*

*Je tiens également à adresser mes sincères remerciements aux membres du jury pour l'intérêt qu'ils ont porté à mon mémoire et pour avoir accepté de le juger.*

*Je remercie tous les enseignants qui ont contribué à ma formation ainsi que le personnel du département de mathématiques .*

Enfin, je remercie du fond du cœur ma famille, mes amis et mes camarades, pour leur soutien tout au long de ce parcours .

**Merci à tous.**

*Hadil Bentrcia*

# Notations et symbols

| | | |
|---|---|---|
| $\mathbb{I}_A(x)$ | : | désigner la fonction indicatrice de l'ensemble $A$. |
| $F_n$ | : | Fonction de répartition empirique. |
| $f_h$ | : | Estimateur par l'histogramme de la densité. |
| $f_n$ | : | Estimateur à noyau de la densité. |
| $\mathbf{E}[\cdot]$ | : | Espérance mathématique. |
| $\mathbf{Var}[\cdot]$ | : | Variance mathématique. |
| $\mathbf{B}[\cdot]$ | : | Biais mathématique. |
| $\mathbf{MSE}[\cdot]$ | : | Erreur quadratique moyenne. |
| $\mathbf{MISE}[\cdot]$ | : | Erreur quadratique emoyenne intégré. |
| $\mathbf{AMSE}[\cdot]$ | : | Erreur quadratique moyenne asymptotique. |
| $\mathbf{AMISE}[\cdot]$ | : | Erreur quadratiqu emoyenne intégré asymptotique |
| $h^*_{opt}$ | : | Paramètre de lissage optimale globale. |
| $h_{opt}$ | : | Paramètre de lissage optimale locale. |
| $Q$ et $Q_n$ | : | Fonctions des quantiles et quantiles empiriques. |
| $\overset{D}{\leadsto}$ | : | Convergence en loi. |
| $(g * f)(x)$ | : | Produit de convolution entre $f$ et $g$. |

# Table des matières

# Table des figures

# Introduction

L'ajustement des modèles de probabilités vise à trouver le modèle statistique optimal reflétant fidèlement la structure réelle des données. Cela inclut le choix du modèle approprié, l'estimation précise de ses paramètres, ainsi que la comparaison de ses résultats avec les données réelles pour évaluer sa qualité et son efficacité. Plus le modèle est bien ajusté, plus il est capable d'expliquer et de prévoir les phénomènes, ce qui en fait un outil d'analyse et de prise de décision plus performant. Comme les modèles statistiques varient en fonction de la nature des données et des phénomènes étudiés, plusieurs méthodes d'ajustement des modèles ont vu le jour. Elles sont généralement classées en deux grandes catégories : les méthodes paramétriques et les méthodes non paramétriques. Les méthodes paramétriques sont des techniques statistiques qui supposent que les données suivent des distributions connues avec des paramètres fixes. Elles sont utilisées pour leur efficacité et leur précision, notamment avec de grands échantillons. Leur objectif est d'estimer la fonction de densité de probabilité ainsi que ses paramètres en utilisant des téchniques comme le maximum de vraisemblance ou les moments. Elles servent également à tester des hypothèses. Malgré leurs avantages, ces méthodes présentent des limites telles que la dépendance à des hypothèses qui peuvent être erronées concernant la forme de la distribution, une faible performance avec des échantillons de petite taille ou hétérogènes, une sensibilité aux valeurs aberrantes,

et un éventuel écart entre les données réelles et les modèles théoriques.C'est pourquoi, on a parfois recours aux méthodes non paramétriques, qui constituent une alternative efficace aux méthodes paramétriques. Elles ne supposent pas une forme spécifique de distribution et ne dépendent pas de paramètres fixes, mais se basent directement sur les données disponibles. Elles s'appuient souvent sur l'ordre ou la fréquence des valeurs plutôt que sur leurs valeurs numériques.

Dans ce travail, nous aborderons deux méthodes paramétriques, dans le but d'étudier comment ajuster des modèles de probabilité et analyser la qualité de l'adéquation entre les modèles théoriques et les données empiriques. Ce mémoire est composé deux chapitres :

**Dans le premier chapitre**, nous aborderons l'estimation de la densité de probabilité, selon deux méthodes : l'histogramme et la méthode du noyau. Nous étudierons également l'estimation des quantiles et quelques-unes de ses propriétés.

**Dans le deuxième chapitre**, nous étudions les tests non paramétriques et leur importance dans la validation et l'ajustement des modèles probabilistes, et ce à travers l'analyse des données par des comparaisons et des déductions basées sur des échantillons. Le but est de vérifier ou de rejeter les hypothèses et de choisir le test optimal en termes d'efficacité et de précision, et ce en se basant sur des travaux antérieurs.

Enfin, nous consacrons dans chaque chapitre une partie de l'étude à l'aspect appliqué, où nous appuyons sur la simulation de données dans le but de comparer les valeurs empiriques extraites des échantillons avec les valeurs théoriques, et d'évaluer la performance des méthodes d'estimation et des tests abordés.

# Chapitre 1

# Estimation non paramétrique

Soit $(X_1, ..., X_n)$ un échantillon d'une variable aléatoire (v.a) $X$ suivant une loi de probabilité inconnue $F(x) := P(X \leq x)$. L'estimation de la fonction de distribution $F$ pose un majeur problème en statistique mathématique. L'une des méthodes utilisées est l'estimation paramétrique qui consiste à choisir un modèle de probabilité paramétré $F_\theta$ adéquat à la distribution $F$. Les méthodes d'adéquation (ou d'ajustement) sont basées sur des tests statistiques tels que le test de kolmogorov, le test de Cramer von-mises, le test de test d'Anderson-Darling.... Cependant le choix d'un modèle approprié de $F$ n'est pas assez facile en général. En effet, les p-valeurs correspondantes aux tests statistiques indiqués sont rarement supérieures à 0.05. Ainsi, la statistique non paramétrique est une approche alternative de la méthode paramétrique. Cette deuxième méthode est basée que sur l'échantillon $(X_1, ..., X_n)$ et qu'elle n'exige aucun modèle probabiliste à priori. L'estimateur non paramétrique usuel de la distribution $F$ est la fonction de répartition empirique :

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq x\}}, \tag{1.1}$$

où $\mathbb{I}_A$ désigne la fonction indicatrice de l'ensemble $A$. D'après le théorème de Glivenko-Cantelli $F_n$ converge uniformément vers $F$ presque surement. L'estimation de la densité de probabilité $f(x) := dF(x)/dx$, à son tour pose aussi un problème en statistique inférentielle. Bien que l'estimateur $F_n$ converge vers $F$ mais ceci ne peut pas servir l'estimation de $f$ et vu son handicap de discontinuité. Dans les prochains paragraphes, nous exposons la célèbre méthode d'estimation non-paramétrique de la densité connue par "l'estimation à noyau".

## 1.1 Estimation de la densité par l'histogramme

Supposons que le support d'une densité de probabilité $f$ est $[0,1]$ et soit $c_k := [\frac{k-1}{m}, \frac{k}{m}[$, $k = 1, ..., m$ une partition de $[0,1]$. Soit :

$$p_k := F\left(\frac{k}{m}\right) - F\left(\frac{k-1}{m}\right) = \int_{c_k} f(x)dx = \int \mathbb{I}_{c_k}(x) f(x)dx = \mathbf{E}\left[\mathbb{I}_{c_k}(X)\right].$$

Observons qu'on a approché $f$ par une fonction en escaliers, constante par morceaux sur l'intervale $c_k$, définie par :

$$f_h(x) := \sum_{k=1}^{m} \frac{p_k}{h} \mathbb{I}_{c_k}(x),$$

où $h := 1/m$, étant la longueur de $c_k$. Ainsi, un estimateur de $f$ peut etre donné en estimant $f_h$, en d'autres termes :

$$\widehat{f}_h(x) := \sum_{k=1}^{m} \frac{\widehat{p}_k}{h} \mathbb{I}_{c_k}(x),$$

où

$$\widehat{p}_k := \widehat{\mathbf{E}}\left[\mathbb{I}_{c_k}(X)\right] = n^{-1} \sum_{i=1}^{n} \mathbb{I}_{c_k}(X_i).$$

Nous avons les propriétés suivantes :

1. $\mathbf{E}\left[\widehat{f}_h\right] = p_k/h$, par conséquent le biais de $\widehat{f}_h$ est $\mathbf{B}(\widehat{f}_h) = p_k/h - f_h$.
2. $\mathbf{Var}\left[\widehat{f}_h\right] = p_k(1-p_k)/\left(nh^2\right)$.
3. $\mathbf{MSE}\left[\widehat{f}_h\right] = \mathbf{B}^2(\widehat{f}_h) + \mathbf{Var}\left[\widehat{f}_h\right]$, ce qui implique que :

$$\mathbf{MSE}(\widehat{f}_h) = \left(\frac{p_k}{h} - f(x)\right)^2 + \frac{p_k(1-p_k)}{nh^2}.$$

4. $\mathbf{MSIE}(\widehat{f}_h) = \int \mathbf{MSE}(\widehat{f}_h(x))dx$, ainsi :

$$\mathbf{MSIE}(\widehat{f}_h) = \int f^2(x)dx + \frac{1}{nh} - \frac{1}{h}\left(1+\frac{1}{n}\right)\sum_{k=1}^{m} p_k^2.$$

5. $h^*_{opt} := \arg\min_h \mathbf{MISE}(\widehat{f}_h)$, ce qui donne :

$$h^*_{opt} = \left(6/\int f^2(x)dx\right)^{\frac{1}{3}} n^{-1/3}.$$

6. La vitesse de convergence de $\widehat{f}_h$ vers $f$ est d'ordre $n^{-2/3}$.

**Remarque 1.1.1** *Les preuves de ces propriétés se trouvent en détail dans [3], [6], [9], [18], [21].*

**Remarque 1.1.2** *Plus généralement, si le support de la densité $f$ est $[a,b]$, on peut prendre $c_k = [a+(k-1)h, a+kh[$ où $h = (b-a)/m$, $k = 1, ..., m-1$.*

## 1.2 Estimation de la densité par la méthode du noyau

Rappelons que la densité de probabilité $f$ est la dérivée de la fonction de répartition $F$, en d'autres termes :

$$f(x) = \frac{dF(x)}{dx} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}.$$

Pour obtenir un estimateur $f_n$ de $f$ , il suffit de remplacer $F$ par la fonction de répartition empirique $F_n$ définie dans.(1.1), c'est- à -dire :

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq x+h\}} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq x-h\}}}{2h},$$

laquelle peut être réécrite en :

$$f_n(x) = \frac{1}{2nh}\left(\sum_{i=1}^{n} \mathbb{I}_{\left\{\frac{X_i - x}{h} \leq 1\right\}} - \sum_{i=1}^{n} \mathbb{I}_{\left\{\frac{X_i - x}{h} \leq -1\right\}}\right) = \frac{1}{2nh}\sum \mathbb{I}_{\left\{-1 \leq \frac{X_i - x}{h} \leq 1\right\}},$$

ainsi :

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2}\mathbb{I}_{\left\{\left|\frac{X_i - x}{h}\right| \leq 1\right\}}.$$

Cette dernière expression peut être réécrite sous la forme suivante :

$$f_{n,\mathbb{K}}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{K}\left(\frac{X_i - x}{h}\right),$$

où $\mathbb{K}(t) := \frac{1}{2}\mathbb{I}_{\{|t|<1\}}$ comme étant une fonction poids (ou noyau, kernel en anglais). Cette formule est connue par l'estimateur à noyau uniforme proposé pour la première fois par Rosenblatt en 1956 [21] . L'estimateur $f_n$ étant discontinu sur

son support, donc pour le rendre lisse, Parzen (1962) [18], propose de choisir une fonction noyau continue $K$ de telle sorte que :

$$f_{n,K}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{1.2}$$

**Proposition 1.2.1** *Si $K$ est une densité ($K > 0, \int_{-\infty}^{\infty} K(t) = 1$) alors $f_{n,K}$ l'est aussi et a les mêmes propriétés de $K$, c'est-à- dire : $f_{n,K}(x)\,dx = 1$, la continuité et la dérivabilité.*

**Preuve 1.2.1** *Puis que $K \geq 0$ alors $f_{n,K} \geq 0$ et que les propriétés de dérivabilité de $K$ se projettent sur l'estimateur $f_{n,K}$. Montrons que $\int_{-\infty}^{\infty} f_{n,K}(x)\,dx = 1$. En effet, nous avons :*

$$\int_{-\infty}^{\infty} f_{n,K}(x)\,dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{X_i - x}{h}\right)dx.$$

*En utilisant le changement de variable $t = (X_i - x)/h$ on obtient :*

$$\begin{aligned}\int_{-\infty}^{\infty} f_{n,K}(x)\,dx &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(t)\,hdt \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(t)\,dt = \int_{-\infty}^{\infty} K(t)\,dt = 1.\end{aligned}$$

Pour la suite, nous utiliserons les notions suivantes :

$$\mu_2(K) := \int t^2 K(t)\,dt,\ R(K) := \int K^2(t)\,dt,\ R(H) := \int H^2(t)\,dt,$$

$$R(f) := \int f^2(x)\,dx \text{ et } R(f'') := \int f''^2(x)\,dx\ .$$

**Proposition 1.2.2** *Soit $f_{n,K}$ l'estimateur de $f$, défini en (1.2), et $K$ un noyau, borné, symétrique, de moment d'ordre 2 fini, vérifiant les propriétés suivantes :*

1. $\mathbf{E}\left[f_{n,K}(x)\right] = f_{n,K}(x) + \left(h^2/2\right) f''(x)\,\mu_2(K) + o\left(h^2\right)$, ce qui implique que :

$$\mathbf{B}\left[f_{n,K}(x)\right] = \left(h^2/2\right) f''(x)\,\mu_2(K) + o\left(h^2\right).$$

2. $\mathbf{Var}\left[f_{n,K}(x)\right] = (1/nh)\, f(x)\, R(K) + o\left(1/(nh)\right)$.

3. $\mathbf{AMSE}\left[f_{n,K}(x)\right] = \left(h^4/4\right) f''^2(x)\,\mu_2^2(K) + (nh)^{-1} f(x)\, R(K) + o\left(1/(nh)\right) + o\left(h^2\right)$.

4. $\mathbf{MSE}\left[f_{n,K}(x)\right] = \left(h^4/4\right) f''^2(x)\,\mu_2^2(K) + (1/nh)\, f(x)\, R(K)$.

5. $\mathbf{AMISE}\left[f_{n,K}(x)\right] = \left(h^4/4\right)\mu_2^2(K)\, R(f'') + (1/nh)\, R(K)$.

6. $h^*_{opt} := \arg\min_h \mathbf{MISE}\left[f_{n,K}(x)\right]$ ce qui signifie que :

$$h^*_{opt} = \left(R(K)/\mu_2^2(K)\, R(f'')\right)^{1/5} n^{-1/5}.$$

7. $h_{opt} := \arg\min \mathbf{MISE}\left[f_{n,K}(x)\right]$ ce qui signifie que :

$$h_{opt} = \left(R(K)\, f(x)/\mu_2^2(K)\, f''(x)^2\right)^{1/5} n^{-1/5}.$$

8. La vitesse de convergence de $f_n$ vers $f$ est d'ordre $n^{-4/5}$.

**Remarque 1.2.1** *Les preuves de ces propriétés aussi se trouvent en détail dans : [3], [6], [16], [18], [21].*

Ce qui conduit à définir un estimateur de la fonction de répartition $F$ par :

$$F_{n,k}(x) := \int_{-\infty}^{x} f_{n,k}(t)\, dt = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x} K\left(\frac{X_i - t}{h}\right) dt.$$

En faisant le changement de variables $s = (X_i - t)/h$, on obtient :

$$F_{n,k}(x) = \frac{1}{n}\sum_{i=1}^{n}\int_{(X_i-x)/h}^{\infty} K(s)\, ds.$$

On pose $H(s) := \int_s^{\infty} K(t)\, dt$, ce qui implique que $-dH(s) = K(s)\, ds$, Par conséquent :

$$F_{n,K}(x) = -\frac{1}{n}\sum_{i=1}^{n}\int_{(X_i-x)/h}^{\infty} dH(s) = -\frac{1}{n}\sum_{i=1}^{n}\left\{H(\infty) - H\left(\frac{X_i - t}{h}\right)\right\}.$$

Comme $H(\infty) = 0$, alors :

$$F_{n,K}(x) = n^{-1}\sum_{i=1}^{n} H\left(\frac{X_i - x}{h}\right). \tag{1.3}$$

**Propriété 1.2.1** *L'estimateur $F_{n,K}$ défini en (1.3) vérifie les propriétés suivantes :*

1. $\mathbf{E}[F_{n,K}(x)] = F(x) + (h^2/2) f'(x) \mu_2(K) + o(h^2)$, ce qui implique :

$$\mathbf{B}[F_{n,K}(x)] = \left(h^2/2\right) f'(x) \mu_2(K) + o\left(h^2\right).$$

2. $\mathbf{Var}[F_{n,K}(x)] = (h/n) f(x) R(H) + o(h/n)$.
3. $\mathbf{AMSE}[F_{n,K}(x)] = (h^4/4) f''(x) \mu_2^2(K) + (h/n) f(x) R(H) + o(h/n) + o(h^2)$.
4. $\mathbf{MSE}[F_{n,K}(x)] = (h^4/4) f''(x) \mu_2^2(K) + (h/n) f(x) R(H)$.
5. $\mathbf{AMISE}[F_{n,K}(x)] = \int \mathbf{AMSE}[F_{n,K}(x)]\, dx$, ce qui signifie que :

$$\mathbf{AMISE}[F_n(x)] = \left(h^4/4\right) \mu_2^2(K) R(f') + (h/4) R(H).$$

6. $h^*_{opt} = \arg\min_h \mathbf{MISE}[F_{n,K}(x)] = \left(R(H)/\mu_2^2(K) R(f')\right)^{1/3} n^{-1/3}$.

7. $h_{opt} = \arg\min_h \mathbf{MISE}\left[F_{n,K}(x)\right] = \left(R(H) f(x) / \mu_2^2(K) f'(x)^2\right)^{1/3} n^{-1/3}$.

**Propriété 1.2.2** *L'estimateur $F_n$ défini en (1.1), vérifie :*

- $\mathbf{B}\left[F_n(x)\right] = 0$.
- $\mathbf{Var}\left[F_n(x)\right] = \frac{1}{n}F(x)(1-F(x))$.
- $F_n$ converge vers $F$ en probabilité, en moyenne quadratique et en loi.
- $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0$ quand $n \to \infty$, presque sûrement.

**Exemple 1.2.1** *Nous allons comparer l'estimation par noyau de quelques lois à leurs densités théoriques :*

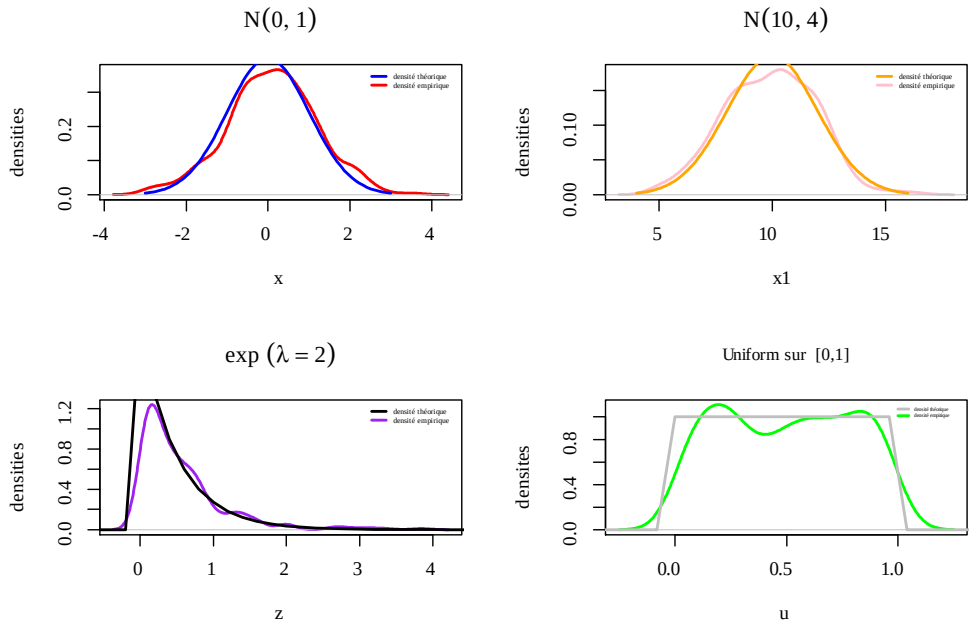


FIG. 1.1 – Comparaison de l'estimation de densité par noyau avec la densité théorique pour les lois normale, exponentielle et uniforme.

### 1.2.1 Quelques noyaux usuels

Les noyaux couramment utilisés sont :

- Les noyaux uniforme, triangulaire et quadratique respectivement :

$$\frac{1}{2}\mathbb{I}_{\{|t|<1\}},\ (1-|t|)\mathbb{I}_{\{|t|<1\}} \text{ et } \frac{15}{16}(1-t^2)^2\mathbb{I}_{\{|t|<1\}}.$$

- Les noyaux d'Epanechnikov, gaussien, cubique, et circulaire respectivement :

$$\frac{3}{4}(1-t^2)\mathbb{I}_{\{|t|<1\}},\ \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}},\ t\in\mathbb{R},\ \frac{35}{32}(1-t^2)^3\mathbb{I}_{\{|t|<1\}} \text{ et } \frac{\pi}{4}\cos(\frac{\pi}{2}t)\mathbb{I}_{\{|t|<1\}}.$$

Nous présentons ci-dessous un graphique illustrant la comparaison entre les noyaux étudiés.

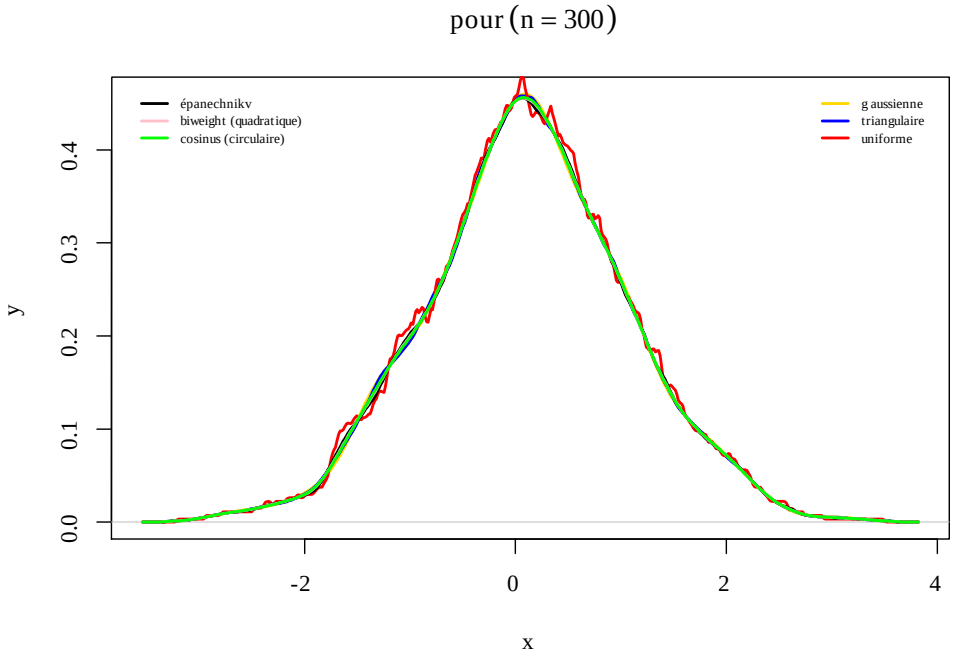


FIG. 1.2 – Quelques noyaux usuels.

### 1.2.2 Efficacité relative

Lors de la définition d'un estimateur à noyau, il faut non seulement choisir la taille de la fenêtre $h$, mais aussi sélectionner le noyau $K$. Pour identifier le noyau optimal, il suffit d'insérer la valeur de la fenêtre optimale $h^*_{opt}$ dans la $\mathbf{AMISE}\left[f_{n,K}\left(x\right)\right]$ qui nous donne :

$$\mathbf{AMISE}\left[f_{n,K}\left(x\right)\right] = \frac{h^{*4}_{opt}}{4}\mu_2^2\left(K\right)\left(f''\right) + \frac{R\left(K\right)}{nh^*_{opt}}.$$

Plus précisément :

$$\mathbf{AMISE}\left[f_{n,K}\left(x\right)\right] = \frac{1}{4}\left(\frac{R\left(K\right)}{\mu_2^2\left(K\right)R\left(f''\right)}\right)^{4/5}\mu_2^2\left(K\right)R\left(f''\right)n^{-4/5} + \left(\frac{R\left(K\right)}{\mu_2^2\left(K\right)R\left(f''\right)}\right)^{-1/5}n^{-4/5}R\left(K\right).$$

On peut montrer que cette dernière formule peut être mise sous la forme suivante :

$$\mathbf{AMISE}\left[f_{n,K}\left(x\right)\right] = \frac{5}{4}\left\{R\left(K\right)^4\mu_2^2\left(K\right)R\left(f''\right)\right\}^{1/5}n^{-4/5}. \tag{1.4}$$

La minimisation de (1.4) par rapport à $K$ donne comme solution le noyau dit d'Epanechnikov. Ainsi on définit l'efficacité relative par :

$$eff(K) = \frac{\mathbf{AMISE}\left(K_{opt}\right)}{\mathbf{AMISE}\left(K\right)} = \left(\frac{R\left(K_{opt}\right)^4\mu_2^2\left(K_{opt}\right)}{R\left(K\right)^4\mu_2^2\left(K\right)}\right)^{1/5} \leq 1,$$

où $K_{opt}$ le noyau d'Epanechnikov.

### 1.2.3 Choix du paramètre de lissage

Le paramètre $h$ dépend des dérivées, quantité $f'$ et $f''$ qui sont elles-mêmes inconnues. Pour résoudre ce problème, on propose une méthode :

**La Règle de référence à la loi normale** une méthode statistique qui repose sur l'hypothèse que les données ou les variables étudiées suivent une distribution normale ce qui permet d'utiliser la densité de loi normale :

$$f(x) = \frac{1}{\delta\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2\right\}.$$

On calcule la dérivée première :

$$f'(x) = \frac{x-\mu}{\delta^3\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2\right\},$$

ainsi que la dérivée deuxième :

$$f''(x) = \frac{\delta - x + \mu}{\delta^4\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2\right\}.$$

Après avoir calculé les dérivées, nous calculons $R(f'')$ comme suit :

$$\begin{aligned}
R(f'') &= \int f''(x)^2 dx = \int \left(\frac{\delta - x + \mu}{\delta^4\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2\right\}\right)^2 dx \\
&= \frac{1}{\delta^8 2\pi} \int (\delta - x + \mu)^2 \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2\right\}^2 dx \\
&= \frac{1}{\delta^6 2\pi} \int \left(1 - \frac{x+\mu}{\delta}\right)^2 \exp\left\{-\left(\frac{x-\mu}{\delta}\right)^2\right\} dx \\
&= \frac{1}{\delta^6 2\pi} \int \left(1 + \left(\frac{x+\mu}{\delta}\right)^2 - 2\left(\frac{x+\mu}{\delta}\right)\right) \exp\left\{-\left(\frac{x-\mu}{\delta}\right)^2\right\} dx.
\end{aligned}$$

En effectuant le changement de variable suivant : $t/2 = x - \mu/\delta \Longrightarrow x = \delta/\sqrt{2}t + \mu \Longrightarrow dx = \delta/\sqrt{2}dt$. Ce qui permet d'obtenir $R(f'')$ comme suit :

$$R(f'') = \frac{1}{\delta^6 2\pi} \int \left(1 - 2t/\sqrt{2} + t^2/2\right) e^{-t^2/2} \delta/\sqrt{2} dt,$$

comme $\mathbf{E}\left[N\left(0,1\right)\right] = 0$ et $\mathbf{Var}\left[N\left(0,1\right)\right] = 1$, par conséquent :

$$R(f'') = \frac{1}{\delta^5 2\sqrt{\pi}} + \frac{1}{\delta^5 4\sqrt{\pi}} = \frac{1}{\delta^5} \frac{3}{8\sqrt{\pi}},$$

ainsi, pour un $R(f'')$ fixé, on trouve :

$$h^*_{opt} = \left(\frac{8\sqrt{\pi} R(K)}{3\mu_2^2(K)}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} \widehat{\delta}.$$

Où $\widehat{\delta}$ est un estimateur avec biais de la variance définie comme :

$$\widehat{\delta} := \left(n^{-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\right)^{-1/2}.$$

### 1.2.4 Estimation à noyau de la fonction des quantiles

Soit $X$ une variable aléatoire ayant une loi de probabilité $F\left(x\right) := P\left(X \leq x\right)$. La fonction des quantiles, ou l'inverse généralisée, associée à $F$ est définie par :

$$Q\left(s\right) := F^{\longleftarrow}\left(s\right) = \inf\left\{x : F\left(x\right) \geq s\right\}, \text{ pour } 0 < s < 1.$$

On dit que :

* $Q(0.5)$ est la médiane .

* $Q(0.25)$ le premier quartile .

* $Q(0.75)$ le troisième quartile.

* $Q(0.1)$ le premier décile .

* $Q(0.8)$ le huitième décile.

**Définition 1.2.1** *Soient $X_{1:n} \leq ... \leq X_{n:n}$ les statistiques d'ordre associées à un échantillon $(X_1, ..., X_n)$ d'une variable aléatoire $X$ ayant une loi de probabilité $F$. La fonction empirique des quantiles est définie par :*

$$Q_n(s) = F_n^{-1}(s) = \inf\{x \in \mathbb{R} : F_n(x) \geq s\}, \text{ pour } 0 < s < 1.$$

*étant $F_n$ la fonction de répartition emprique usuelle. On montre que :*

$$Q_n(s) = X_{i:n}, \text{ pour } \frac{i-1}{n} < s \leq \frac{i}{n}, \ i = 1, 2, ..., n.$$

*Observons que pour tout $0 < p < 1$ on a :*

$$\frac{[np]}{n} < p \leq \frac{[np]+1}{n}.$$

*Donc si on prend $i = [np] + 1$, on en déduit que $Q_n(p) = X_{[np]+1:n}$, ce qui fournit un estimateur non paramétrique du quantile d'ordre $p$.*

Observons maintenant que :

$$\begin{aligned} X_{[np]+1:n} &= n \int_{\frac{[np]}{n}}^{\frac{[np]+1}{n}} Q_n(t)dt = n \int_0^1 Q_n(t)\mathbb{I}_{\left(\frac{[np]}{n} < t \leq \frac{[np]+1}{n}\right)}dt \\ &= n \int_0^1 Q_n(t)\mathbb{I}_{\left(\frac{[np]}{n} - p < t - p \leq \frac{[np]+1}{n} - p\right)}dt. \end{aligned}$$

Rappelons que $[x] \leq x \leq [x]+1$, par conséquent :

$$\begin{aligned} X_{[np]+1,n} &= n\int_0^1 Q_n(t)\mathbb{I}_{\left(\frac{np-1}{n}-p\,<\,t-p\,\leq\frac{np+1}{n}-p\right)}dt \\ &= n\int_0^1 Q_n(t)\mathbb{I}_{\left(-\frac{1}{n}<\,t-p\,\leq\frac{1}{n}\right)}dt \\ &= n^{-1}\int_0^1 Q_n(t)\mathbb{I}_{\left(\left|\frac{p-t}{n^{-1}}\right|\leq 1\right)}dt. \end{aligned}$$

En d'autres termes :

$$X_{[np]+1:n} = \frac{1}{h_0}\int_0^1 Q_n(t)k^*\left(\frac{p-t}{h_0}\right)dt,$$

où $k^*(x) = \mathbb{I}_{|x|\leq 1}$ et $h_0 := 1/n$. Notons que $\int_{-\infty}^{+\infty} k^*(x)\,dx = 1$. Ansi on a définit un estimateur à noyau des quantiles d'ordre $p$ comme étant le produit de convolution entre $Q_n$ et $K_h$ :

$$Q_{n,k}(p) = (Q_n * k_h)(p) = \int_0^1 Q_n(t)k_h(p-t)\,dt,$$

où $k_h(x) := \frac{1}{h}k\left(\frac{x}{h}\right)$. D'après la représentation on peut écrire :

$$\begin{aligned} Q_{n,k}(p) &= \sum_{i=1}^n X_{i:n}\int_{(i-1)/n}^{i/n} h^{-1}k\left(\frac{p-t}{h}\right)dt \\ &= \sum_{i=1}^n X_{i:n}h^{-1}\left[K\left(\frac{p-i/n}{h}\right) - K\left(\frac{p-(i-1)/n}{h}\right)\right], \end{aligned}$$

où $K(t) := \int_{-\infty}^t k(s)ds$.

**Propriété 1.2.3** *Sous certaines conditions sur le noyau $K$ et le paramètre de lissage $h$ on a :*

1. Le biais de $Q_{n,k}(p)$ est :

$$\mathbf{B}\left[Q_{n,k}\left(x\right)\right]=\frac{1}{2}h^{2}\left[\int_{0}^{1}y^{2}k(y)dy\right]Q^{''}(p)+o(h^{2})+O\left(n^{-1}\right).$$

2. Par des arguments similaires, on obtient :

$$\mathbf{Var}\left[Q_{n,k}\left(x\right)\right]=\frac{p(1-p)}{n}\left(Q_{n}^{\prime}\left(p\right)\right)^{2}-(h/n)\left(Q_{n}^{\prime}\left(p\right)\right)^{2}\int yk(y)K(y)dy+o(h/n).$$

Les formules asymptotiques de $\mathbf{B}\left[Q_{n,k}\left(x\right)\right]$ et de $\mathbf{Var}\left[Q_{n}\left(x\right)\right]$ sont obtenues par Sheather et Marron (1990). [23].

3. Erreur quadratique moyenne de $Q_{n,k}\left(x\right)$ :

Supposant que $Q''$ est continue au voisinage de $p$ et que $k$ est symétrique autour de 0. Pour tout $p\in(0,1)$ et pour $F$ est symétrique et $p\neq 0$ on a :

$$\begin{aligned}\mathbf{MSE}\left[Q_{n.k}\left(p\right)\right]&=\frac{p(1-p)}{n}\left(Q^{\prime}\left(p\right)\right)^{2}+\frac{h^{4}}{4}\left(Q^{\prime\prime}\left(p\right)\right)^{2}\mu_{2}^{2}(k)\\&\quad-\frac{h}{n}\left(Q^{\prime}\left(p\right)\right)^{2}\varphi(k)+o\left(h/n+h^{4}\right).\end{aligned}$$

Pour plus de détails voir Falk (1984),[11].

**Preuve 1.2.2** 1.*Le biais de* $Q_{n,k}(p)$ *est :*

$$\mathbf{B}\left[Q_{n,k}(p)\right]=\mathbf{E}\left[Q_{n,k}(p)-Q\left(p\right)\right]=\mathbf{E}\left[Q_{n,k}(p)\right]-Q\left(p\right).$$

*Nous avons*

$$\begin{aligned}\mathbf{E}\left[Q_{n,k}(p)\right]&=\mathbf{E}\left[\int_{0}^{1}Q_{n}\left(x\right)h^{-1}k\left(\frac{p-x}{h}\right)dx\right]\\&=\sum_{i=1}^{n}\mathbf{E}\left[X_{i:n}\right]\int_{(i-1)/n}^{i/n}h^{-1}k\left(\frac{p-x}{h}\right)dx.\end{aligned}$$

*D'après la formule* (4.6.3) *dans David and Nagaraja* (2003) *[10], on a :*

$$\mathbf{E}\left[X_{i:n}\right] = Q\left(\frac{i}{n+1}\right) + O\left(n^{-1}\right), \text{ quand } n \to \infty,$$

*ce qui implique que :*

$$\begin{aligned}\mathbf{E}\left[Q_{n,k}(p)\right] = \sum_{i=1}^{n} Q\left(\frac{i}{n+1}\right) \int_{(i-1)/n}^{i/n} h^{-1} k\left(\frac{p-x}{h}\right) dx \\ + O\left(\frac{1}{n}\right) \int_{0}^{1} h^{-1} k\left(\frac{p-x}{h}\right) dx.\end{aligned}$$

*Comme* $0 < \int_0^1 h^{-1} k\left(\frac{p-x}{h}\right) dx < 1$, *donc :*

$$\mathbf{E}\left[Q_{n,k}(p)\right] = \sum_{i=1}^{n} Q\left(\frac{i}{n+1}\right) \int_{(i-1)/n}^{i/n} h^{-1} k\left(\frac{p-x}{h}\right) dx + O\left(n^{-1}\right).$$

*En utilisant le théorème des accroissements finis, on peut écrire :*

$$Q\left(x\right) - Q\left(\frac{i}{n+1}\right) = \left(t - \frac{i}{n+1}\right) Q'\left(p_i\right),$$

*où* $\frac{i}{n+1} < p_i < x < \frac{i}{n}$. *On note que* $Q'\left(p_i\right)$ *est finie et que pour* $\frac{i-1}{n} < x < \frac{i}{n}$ *on a :*

$$\left| x - \frac{i}{n+1} \right| < \frac{1}{n}.$$

*Par conséquent :*

$$Q\left(x\right) = Q\left(\frac{i}{n+1}\right) = O\left(n^{-1}\right),$$

*uniformément sur* $t$. *Ainsi :*

$$\mathbf{E}\left[Q_{n,k}(p)\right] = \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} Q\left(x\right) h^{-1} k\left(\frac{p-x}{h}\right) dx + O\left(n^{-1}\right).$$

*Observons maintenant que :*

$$\begin{aligned}\mathbf{B}\left[Q_{n,k}(p)\right] &= \sum_{i=1}^{n}\int_{(i-1)/n}^{i/n}\left(Q\left(x\right)-Q\left(p\right)\right)h^{-1}k\left(\frac{p-x}{h}\right)dx+O\left(n^{-1}\right)\\ &= \int_{0}^{1}\left(Q\left(x\right)-Q\left(p\right)\right)h^{-1}k\left(\frac{p-x}{h}\right)dx+O\left(n^{-1}\right).\end{aligned}$$

*En utilisant le changement de variables on a* $y=(p-x)/h$ *on obtient* $x=p-yh$ *et* $dx=-hdy$. *Donc si* $x=0\Longrightarrow y=p/h$ *et si* $x=1\Longrightarrow y=p-1/h$, *par conséquent, le biais de* $Q_{n,k}$ *devient :*

$$\mathbf{B}\left[Q_{n,k}\left(x\right)\right]=-\int_{p/h}^{p-1/h}k(y)\left(Q\left(yh+p\right)-Q\left(p\right)\right)dy+O\left(n^{-1}\right).$$

*En appliquant la formule de Taylor d'ordre* 2, *on obtient :*

$$Q\left(yh+p\right)=Q\left(p\right)-\left(yh\right)Q^{'}(p)+\left(\left(yh\right)^{2}/2\right)Q^{''}(p)+o(h^{2}),$$

*quand* $n\rightarrow\infty$, *ainsi :*

$$\mathbf{B}\left[Q_{n,k}\left(x\right)\right]=\frac{1}{2}h^{2}\left[\int_{0}^{1}y^{2}k(y)dy\right]Q^{''}(p)+o(h^{2})+O\left(n^{-1}\right).$$

**Exemple 1.2.2** *Nous allons comparer l'estimation par noyau de quelques lois à leurs quantiles théoriques :*

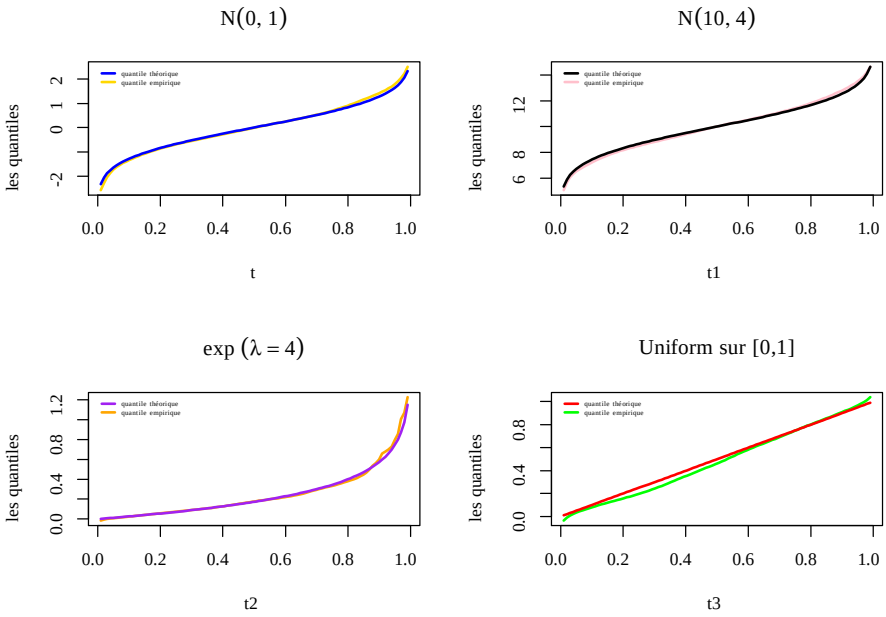


FIG. 1.3 – Comparaison de l'estimation des quantiles par noyau avec les quantiles théoriques pour les lois normale, exponentielle et uniforme.

# Chapitre 2

# Tests non paramétrique

Les tests statistiques sont des outils utilisés pour prendre une décision entre deux hypothèses ($H_0$ et $H_1$), en se basant sur des données extraites d'un échantillon donné. Lors de l'analyse des données de l'échantillon, on ne se repose pas sur une distribution spécifique ou conditionnelle, ce qui rend les tests plus flexibles. C'est ce qu'on appelle les tests non paramétriques qui sont considérés comme plus robustes. Ils sont adaptés aux petits échantillons et sont moins sensibles aux valeurs aberrantes. Ils sont également utilisés avec des données ordinales ou qualitatives.

On distingue deux types de tests : les tests basés sur les rangs et ceux basés sur les distributions. Dans la section qui suit, nous présenterons chaque type à l'aide d'exemples."

## 2.1 Tests basés sur les rangs

Étant donné deux variables aléatoires indépendantes $X$ et $Y$ qui suivent les lois de probabilités $F_X(x) := P(X \leq x)$ et $F_Y(x) := P(Y \leq x)$ respectivement. On

s'intéresse au test d'homogénéité suivant :

$$H_0 : F_X = F_Y \text{ contre } H_1 : F_X \neq F_Y.$$

Soient $(X_1, ..., X_{n_1})$ et $(Y_1, ..., Y_{n_1})$ deux échantillons de $X$ et $Y$ de tailles $n_1$ et $n_2$ respectivement. On considère l'échantillon $(X_1, ..., X_{n_1}; Y_1, ..., Y_{n_2})$ qu'on note par $(Z_1, ..., Z_n)$, où $n := n_1 + n_2$. Ce dernier peut être considéré comme étant un échantillon issu d'une variable aléatoire $Z$ de loi de probabilité $H(z) := P(Z \leq z)$. On définit le rang de l'observation $X_j$ dans l'echantillon $(Z_1, ..., Z_n)$ par :

$$R_n(X_j) := \sum_{i=1}^{n} \mathbb{I}_{(Z_i \leq X_j)} = nH_n(X_j),$$

où $H_n(z) := n^{-1} \sum_{i=1}^{n} \mathbb{I}_{\{Z_i \leq z\}}$ désigne la fonction de répartition empirique associée à $(Z_1, ..., Z_n)$. En d'autres termes, le rang de $X_j$ n'est autre que la position de celle-ci dans l'échantillon ordonné $Z_{1:n} \leq ... \leq Z_{n:n}$. Il est important de souligner qu'on cas d'ex aequo on prend la moyenne des rangs des deux observations.

**Exemple 2.1.1** *Soient* $\{-1, 11, 5\}$ *et* $\{0, 7\}$ *deux échantillons indépendants issuent de deux variables aléatoires* $X$ *et* $Y$ *respectivement. Ici* $n_1 = 3$, $n_2 = 2$, $n = 3 + 2 = 5$ *et*

$$(Z_1, Z_2, Z_3, Z_4, Z_5) = (-1, 11, 5, 0, 7).$$

*Par exemple le rang de* $X_3 \equiv Z_3 = 5$ *est :*

$$\sum_{i=1}^{n} \mathbb{I}_{(Z_i \leq 5)} = \begin{cases} 1 & si \quad Z_i \leq 5 \\ 0 & si \quad Z_i > 5 \end{cases} = 3.$$

Nous allons annoncer quelques tests statistiques qui répondent au problème ci-dessus. Il s'agit de tests statistiques non paramétriques utilisés pour comparer deux échantillons ou plus, afin de déterminer s'il existe des différences statistiquement significatives entre eux. Ces tests se basent sur les statistiques des rangs plutôt que sur l'échantillon lui même. Ceci les rendent particulièrement adaptés lorsque les conditions d'application des tests paramétriques (comme la normalité des distributions) ne sont pas remplies.

### 2.1.1 Test de Wilcoxon

La statistique du test de Wilcoxon est définie par :

$$W(X,Y) := \begin{cases} \sum_{i=1}^{n_1} R(X_i) & \text{si } n_1 \leq n_2, \\ \sum_{i=1}^{n_2} R(Y_i) & \text{si } n_1 \geq n_2, \end{cases}$$

où $R(X_i)$ et $R(Y_i)$ sont les rangs des deux l'observations $X_i$ et $Y_i$ dans l'echantillon $(X_1, ..., X_{n_1}; Y_1, ..., Y_{n_2})$, respectivement. On accepte $H_0$ si $W \leq W_\alpha$ où $W_\alpha$ est le quantile à la statistique de Wilcoxon d'ordre $\alpha$.

**Théoréme 2.1.1** *Si* $\min(n_1, n_2) \geq 10$, *nous utilisons l'approximation de la loi normale :*

$$\mathcal{Z} := \frac{W - \mathbf{E}(W)}{\sqrt{\mathbf{Var}(W)}} \overset{D}{\simeq} N(0,1).$$

*On accepte* $H_0$ *si* $|\mathcal{Z}(0,1)| \leq z_{1-\alpha/2}$ *où* $z_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de loi normal centrée-réduite. Il est important de souligner que si* $n_1 \leq n_2$ :

$$\mathbf{E}(W) = (n_1(n+1))/2 \text{ et } \mathbf{Var}(W) = (n_1 n_2(n+1))/12,$$

*et si* $n_1 \leq n_2$ :

$$\mathbf{E}(W) = (n_2(n+1))/2 \text{ et } \mathbf{Var}(W) = (n_1 n_2(n+1))/12.$$

**Exemple 2.1.2** *Les données sur le taux hebdomadaire de production d'articles de deux lignes de production différentes pour* 11 *semaines sont les suivantes :*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Ligne 1* | 36 | 36 | 38 | 40 | 40 | 41 | 41 | 41 | 42 | 42 | 43 |
| *Ligne 2* | 28 | 29 | 29 | 34 | 37 | 37 | 39 | 40 | 41 | 43 | 44 |

*Pour calculer la statistique de Wilcxon on classe les valeurs des observations par ordre croissant :*

$$28, 29, 29, 34, 36, 36, 37, 37, 38, 39, 40, 40, 40, 41, 41, 41, 41, 42, 42, 43, 43, 44.$$

*Les deux lignes de production ont le même nombre de tailles, ainsi nous choisirons les rangs des observations par ordre croissant uniquement celles de la ligne 1 :*

$$\{R(X_i)\}_{i=1,11} = \{5.5, 5.5, 9, 12, 12, 15.5, 15.5, 15.5, 18.5, 18.5, 20.5\}.$$

*Donc la statistique de Wilcoxon est :*

$$W = \sum_{i=1}^{11} R(X_i) = 148,$$

*et on a :*

$$\mathbf{E}(W) = \frac{11(22+1)}{2} = 126.5 \text{ et } \mathbf{Var}(W) = \frac{11 \times 11(22+1)}{12} = 231.$$

*Ce qui implique que :*

$$\mathcal{Z} = \frac{148 - 126,5}{\sqrt{231}} = 1,41.$$

*Pour* $\alpha = 5\%$ *ona* $\mathcal{Z} = 1,41 \leq z_{1-\frac{\alpha}{2}} = 1,96$ *donc, on accepte* $H_0 : F_X = F_Y$. *En d'autres termes les deux variables aléatoires Ligne 1 et Ligne 2 sont égales en distribution.*

**Code R :**

```
wilcoxon_w_pvaleur = function(x, y) {
combined = c(x,y)
ranks =rank(combined)
n1=length(x)
n2=length(y)
ranks_x = ranks[1 :n1]
ranks_y = ranks[n1+1 :n1+n2]
if (n1 <= n2) {
W = sum(ranks_x)
} else {
W = sum(ranks_y) }
p-value =1-pnorm(W, mean = (n1 * (n1 + n2 + 1)) / 2, sd = sqrt(n1 * n2 * (n1
+ n2 + 1) / 12))
return(list(W = W, p-value = p-value))
}
x = c(36,36,38,40,40,41,41,41,42,42,43)
y = c(28,29,29,34,37,37,39,40,41,43,44)
```

wilcoxon_stat(x,y)

wilcoxon_w_pvaleur(x, y)

Après l'excusions,voici les résultas obtenus :

**w = 148 , p-value = 0.07900476.**

On a la p-valeur est supérieure à 0.05, donc on accepte $H_0$.

### 2.1.2 Test de Mann-Whitney

La statistique du test de Mann-Whitney est définie par :

$$U := \min\left(U_X, U_Y\right),$$

où

$$U_X := W_X - \frac{n_1\left(n_1+1\right)}{2} \text{ et } U_Y := W_Y - \frac{n_2\left(n_2+1\right)}{2},$$

et $W_X := W$, si $n_1 < n_2$ et $W_Y := W$ si $n_1 < n_2$. On accepte $H_0$ si $U \leq U_\alpha$ où $U_\alpha$ est le quantile de Mann-Whitney d'ordre $\alpha$.

**Théoréme 2.1.2** *Si* $\min\left(n_1, n_2\right) \geq 10$ *nous utilisons l'approximation :*

$$\mathcal{Z} = \frac{W_X - \mathbf{E}\left(U\right)}{\sqrt{\mathbf{Var}\left(U\right)}} \overset{D}{\simeq} N(0,1).$$

*On accepte* $H_0$ *si* $|\mathcal{Z}| \leq z_{1-\frac{\alpha}{2}}$ *où* $z_{1-\frac{\alpha}{2}}$ *est le quantile d'ordre* $1-\alpha/2$ *de loi normal centrée réduite, telles que*

$$\mathbf{E}\left(U\right) = \left(n_1 n_2\right)/2, \text{ et } \mathbf{Var}\left(U\right) = \left(n_1 n_2\left(n_1+1\right)\right)/12.$$

**Exemple 2.1.3** *Pour comparer des résultats de deux groupes suivants :*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Group* 1 | 7 | 18 | 9 | 9 | 18 | 12 | 16 | 13 | 10 | 19 | 19 |
| *Group* 2 | 10 | 14 | 7 | 8 | 9 | 17 | 15 | 13 | 15 | 18 | |

*On calcule la statistique de Wilcoxon pour le groupe* 1 *et le groupe* 2 *:*

$$R(X_i) = \{1.5, 5, 5, 7.5, 9, 10.5, 15, 18, 18, 20.5, 20.5\} \Longrightarrow W_X = 130.5$$

*et*

$$R(Y_i) = \{1.5, 3, 5, 7.5, 10.5, 12, 13.5, 13.5, 16, 18\} \Longrightarrow W_Y = 100.5$$

*D'autre part, on a*

$$U_X = 130.5 - (11(11+1))/2 \Longrightarrow U_X = 64.5$$

*et*

$$U_Y = 100.5 - (10(10+1))/2 \Longrightarrow U_Y = 45.5$$

*Donc la valeur observée de la statistique du test de Mann-Whitney est* $U = \min(U_X, U_Y) = U_Y = 45.5$. *En outre, nous avons :*

$$\mathbf{E}(U) = (11 \times 10)/2 = 55, \text{ et } \mathbf{Var}(U) = (11 \times 10(11+1))/12 = 110,$$

*ainsi* $Z = (45.5 - 55)/\sqrt{110} = -0.90$. *Donc* $|\mathcal{Z}| = 0.90 \leq z_{1-\alpha.2} = 1.96$ *alors les deux groupes ont la même distribution.*

**Code R :**

lin1=c(10, 14, 7, 8, 9, 17, 15, 13, 15, 18)

lin2=c(7, 18, 9, 9, 18, 12, 16, 13, 10, 19, 19)

wilcox.test(lin1, lin2)

Aprés l'excusions,voici les résultas obtenus :

**W = 45.5, p-value = 0.5245.**

On a la p-valeur est supérieure à 0.05, donc on accepte $H_0 : F_X = F_Y$.

### 2.1.3 Test de Mood de la médiane

Soient $(X_1, ..., X_{n_1})$ et $(Y_1, ..., Y_{n_2})$ deux échantillons indépendants issus de deux variables aléatoires continues indépendantes $X$ et $Y$ de lois $F$ et $G$ respectivement. On s'interèsse à tester si $X$ et $Y$ ont la même distribution. La statistique de la médiane est définie comme la somme des rangs des observations de la variable $X$ qui sont strictement supérieures à la médiane du couple $(X, Y)$. Plus précisement, la médiane est définie par :

$$M := \begin{cases} Z_{\left[\frac{k+1}{2}\right]:k} & \text{si } k \text{ est impaire,} \\ \\ \frac{Z_{\left[\frac{k}{2}\right]:k} + Z_{\left[\frac{k}{2}+1\right]:k}}{2} & \text{si } k \text{ est paire,} \end{cases}$$

où $Z_{1:k} \leq ... \leq Z_{k:n}$ les statistiques d'ordres assoicées au mélange des deux échantillons $(X_1, ..., X_{n_1})$ et $(Y_1, ..., Y_{n_2})$ noté par $(Z_1, ..., Z_k) := (X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2})$, où $k := n_1 + n_2$. Condidérons le tableau de contingence $(2 \times 2)$ suivant :

| | Échantillon 1 | Échantillon 2 |
|---|---|---|
| Nbr. observations $\geq M$ | $a$ | $b$ |
| Nbr. observations $< M$ | $c$ | $d$ |
| Total | $n_1$ | $n_2$ |

Nous distinguons deux situations :

– Si $\max(n_1, n_2) < 10$ : sous l'hypothèse nulle $H_0 : F = G$, la loi jointe de $(m_1, m_2)$ est la distribution hypergéométrique dans la masse de probabilité est :

$$p = \frac{\binom{n_1}{a}\binom{n_2}{b}}{\binom{k}{a+b}},$$

où $\binom{n}{m} := \frac{n!}{m!(n-m)!}$. Si $p < \alpha = 0.05$, on rejette $H_0$.

– Si $\max(n_1, n_2) > 10$ et $\max(a, b, c, d) > 5$ : on utilise de la statistique de khi-deux :

$$T := \frac{k(ad - bc)^2}{(a+b)(c+d)(a+c)b+d}.$$

Si $\max(n_1, n_2) > 10$ et $\max(a, b, c, d) \leq 5$ : on applique la correction de continuité de Yate :

$$T := \frac{k(|ad - bc| - k/2)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

On rejette $H_0$ si $T > \chi^2_d(\alpha)$, où $\chi^2_d(\alpha)$ est le quantile d'ordre $\alpha = 0.05$ de la loi de khi-deux à $d := (2-1)(2-1) = 1$ degré de liberté. On note que $\chi^2_1(0.05) = 3.84$.

**Exemple 2.1.4** *Il s'agit de comparer les deux population suivantes :*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $X$ | 30 | 140 | 228 | 60 | 79 | 52 | | |
| $Y$ | 99 | 450 | 103 | 120 | 100 | 620 | 703 | 150 |

*Nous avons ici* $n_1 = 6$, $n_2 = 8$, $k = n_1 + n_2 = 14$. *Pour calculer la statistique de la médiane on classe les valeurs des observations par ordre croissant :*

$$30, 52, 60, 79, 99, 100, 103, 120, 140, 150, 228, 450, 620, 703$$

*Comme* $k = 14$ *est paire, la médiane de cet échantillon égale*

$$M = \frac{Z_{\left[\frac{14}{2}\right]:14} + Z_{\left[\frac{14+1}{2}\right]:14}}{2} = \frac{Z_{7:14} + Z_{8:14}}{2} = \frac{103 + 120}{2} = 111.5.$$

*La table de contingence (TC) associée :*

| | *Échantillon 1* | *Échantillon 2* |
|---|---|---|
| *Nbr. observations* $\geq M$ | 2 | 5 |
| *Nbr. observations* $< M$ | 4 | 3 |
| *Total* | 6 | 8 |

*Comme* $\max(6, 8) < 10$, *donc on calcule la statistique*

$$p = \frac{\binom{n_1}{a}\binom{n_2}{b}}{\binom{k}{a+b}} = \frac{\binom{6}{2}\binom{8}{5}}{\binom{14}{7}} = 0.244\,76 > \alpha = 0.05,$$

*donc* $X$ *et* $Y$ *proviennent de la même population.*

**Code R**

X = c(30,140,228,60,79,52) ; n1 = length(X)

Y = c(99,450,103,120,100,620,703,150) ; n2= length(Y)

Z = c(X,Y) ; k = length(Z)

M = median(Z)

a = length(X[X>=M]) ; c = n1-a

b = length(Y[Y>=M]) ; d = n2-b

comb = function(n, x) {

factorial(n) / factorial(n-x) / factorial(x)}

p = comb(n1,a)*comb(n2,b)/comb(k,a+b)

p-value = 0 ; t = a + b ; min-a = max(0, t - n2)

max-a = min(n1, t) ; for (i in min-a :max-a) {

j = t - i ; p-ij = comb(n1, i) * comb(n2, j) / comb(k, t)

if (p-ij =p) {p-value = p-value + p-ij }

}

Donc,les résultat sont :

**p = 0.24475, p-value =0.592074.**

On a la p-valeur est supérieure à 0.05, donc on accepte $H_0 : F_X = F_Y$.

**Exemple 2.1.5** *Il s'agit de comparer les deux population suivantes :*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 15 | 17 | 11 | 13 | 32 | 27 | 14 | 6 | 22 | 33 | 18 |
| $Y$ | 11 | 42 | 62 | 13 | 5 | 24 | 10 | 33 | 18 | 19 | 22 | 16 |

*Nous avons ici* $n_1 = 11$, $n_2 = 12$, $k = n_1 + n_2 = 23$. *Pour calculer la statistique de la médiane on classe les valeurs des observations par ordre croissant. Pour calculer la statistique de la médiane on classe les valeurs des observations par ordre croissant :*

5, 6, 10, 11, 11, 13, 13, 14, 15, 16, 17, 18, 18, 19, 22, 22, 24, 27, 32, 33, 33, 42, 62.

Comme $k = 23$ est impair, la médiane de cet échantillon est égale à :

$$M = Z_{\left[\frac{23+1}{2}\right]:23} = Z_{12:23} = 18.$$

La table de contingence (TC) associée :

| | Échantillon 1 | Échantillon 2 |
|---|---|---|
| Nbr. observations $\geq M$ | 5 | 7 |
| Nbr. observations $< M$ | 6 | 5 |
| Total | 11 | 12 |

Comme $\max(11, 12) > 10$ et $\max(a, b, c, d) \leq 5$, on applique la correction de continuité de Yate :

$$\chi^2_{obs} := \frac{k\left(|ad - bc| - k/2\right)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{23\left(|5 \times 5 - 7 \times 6| - 23/2\right)^2}{(5+7)(6+5)(5+5)(7+5)} = 4.392\,4 \times 10^{-2}.$$

Comme $\chi^2_{obs} = 4.392\,4 \times 10^{-2} < \chi^2_1(0.05) = 3.84$, donc $X$ et $Y$ proviennent de la même population.

**Code R :**

X = c(15,17,11,13,32,27,14,6,22,33,18) ; n1 = length(X)

Y = c(11,42,62,13,5,24,10,33,18,19,22,16) ; n2 = length(Y)

k = n1+n2

Z = c(X,Y)

M =median(Z)

a = length(X[X>=M]) ; c = n1-a

b = length(Y[Y>=M]) ; d = n2-b

TC = matrix(c(a,b,c,d),ncol=2,byrow=TRUE)

Khi2 = chisq.test(TC)

Les résultats obtenus sont

**X-squared = 0.039931.**

### 2.1.4 Test de rangs signés de Wilcoxon

Ce test permet d'étudier la comparaison d'échantillons appariés, et il est défini par :

$$W^+ := \sum_{i=1}^{n_1} R_i \mathbb{I}_{\{|D_i|>0\}},$$

où $D_i$ désigne la différence entre deux observations appariées correspondant à la même unité statistique (individu, objet,...), mesurées dans deux conditions différentes et $R_i = rang(|D_i|)$. On accepte $H_0$ si $W^+ \leq W_\alpha$ tell que $W_\alpha$ est le quantile de Wilcoxon d'ordre $\alpha$.

**Théoréme 2.1.3** *Si* $\min(n_1, n_2) \geq 10$, *nous utilisons l'approximation de la loi normale :*

$$\mathcal{Z} := \frac{W^+ - \mathbf{E}(W^+)}{\sqrt{\mathbf{Var}(W^+)}} \simeq \mathcal{N}(0.1).$$

*Il est important de souligner que*

$$\mathbf{E}(W^+) = (n_1(n_2+1))/2 \text{ et } \mathbf{Var}(W^+) = (n_1(n_2+1)(2n_1+1))/24.$$

*On accepte* $H_0$ *si* $|Z| \leq z_{1-\alpha/2}$ *où* $z_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de loi normal centrée-réduite.*

**Remarque 2.1.1** *:*

**1)** $W^+ - W^- = (n_1(n_2+1))/2$.

**2)** $0 \leq W^+ \leq (n_1(n_2+1))/2$.

**3)** $W^+ = \begin{cases} (n_1(n_2+1))/2 & \text{si} \quad D_i > 0. \\ 0 & \text{si} \quad D_i < 0. \end{cases}$

**Exemple 2.1.6** *Un chercheur veut tester si un nouveau médicament diminue la pression artérielle, alors il mesure la pression artérielle de* 10 *patients avant et après le traitement :*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *avant* | 150 | 160 | 166 | 157 | 170 | 146 | 175 | 158 | 180 | 174 | 190 |
| *après* | 140 | 155 | 150 | 144 | 182 | 150 | 165 | 151 | 175 | 160 | 186 |

*D'abord, on calcule la différence entre les valeurs avant et après, c'est-à-dire* $D_i = X_i - Y_i$*, puis on attribue les rangs* $r_i$ *aux valeurs de* $D_i$ :

| *patients* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $D_i$ | 10 | 5 | 16 | 13 | $-12$ | $-4$ | 10 | 7 | 5 | 14 | 4 |
| $\|D_i\|$ | 10 | 5 | 16 | 13 | 12 | 4 | 10 | 7 | 5 | 14 | 4 |
| $r_i$ | 6.5 | 3.5 | 11 | 9 | 8 | 1.5 | 6.5 | 5 | 3.5 | 10 | 1.5 |

*Donc, la valure de statistique est la somme des rangs des* $D_i$ *positifs :*

$$W^+ = 6.5 + 3.5 + 11 + 9 + 6.5 + 5 + 3.5 + 10 + 1.5 = 56.5.$$

*Au niveau de signification de* 5%*, d'après le tableau de Wilcoxon, on a* $W_\alpha = 11$*. Alors* $W^+ \geq W_\alpha$ *ce qui signifie que le nouveau médicament ne diminue pas la pression artérielle.*

**Code R :**

avant=c(150, 160, 166, 157, 170, 146, 175, 158, 180, 174, 190)

apres=c(140, 155, 150, 144, 182, 150, 165, 151, 175, 160, 186)

wilcox.test(avant,apres,paired =TRUE)

Suite à l'exclusion, voici les résultats obtenus :

**W = 56.5, p-value = 0.04056.**

On a la p-valeur est inférieure à 0.05, donc on rejette $H_0 : F_X = F_Y$.

### 2.1.5 Test de Kruskal-Wallis

Ce test permet de déterminer s'il existe une différence significative entre plusieurs groupes indépendants. L'hypothèse nulle est formulée comme suit $H_0 : F_1 = ... = F_m$. Donc, on définit la statistique comme suit :

$$H := \frac{12}{n(n+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(n+1).$$

Telle que :

- $n$ : Le nombre total d'observations dans tous les groupes.
- $m$ : Le nombre de groupes.
- $n_i$ : Le nombre d'observations du $i$-ème groupe.
- $R_i$ : La somme des rangs .

On accepte l'hypothèse $H_0$ si $H \leq \chi^2_{(m-1)}$ où $\chi^2_{(m-1)}$ est le quantile de la statistique du khi-deux à $(m-1)$ degrès de liberté.

**Exemple 2.1.7** *Nous avons mesuré le temps de réaction de trois groupes et nous souhaitons savoir s'il existe une différence entre ces groupes :*

| | | | | |
|---|---|---|---|---|
| *Group A* | 34 | 36 | 41 | 43 |
| *Group B* | 44 | 37 | 45 | 33 |
| *Group C* | 35 | 39 | 42 | 46 |

*Premièrement, on range les observations comme suit :*

$$33 \quad 34 \quad 35 \quad 36 \quad 37 \quad 39 \quad 41 \quad 42 \quad 43 \quad 44 \quad 45 \quad 46$$

*Après on calcule la somme des rangs* $R_i$, $i = 1, 2, 3$ :

$$\begin{array}{llllll} \textit{Rang } A & 2 & 4 & 7 & 9 & \\ \textit{Rang } B & 10 & 5 & 11 & 1 & \Longrightarrow \\ \textit{Rang } C & 3 & 6 & 8 & 12 & \end{array} \begin{cases} R_1 = & 22 \\ R_2 = & 27 \\ R_3 = & 29 \end{cases}$$

*Nous* $n = 12$ , $n_i = 4$, *pour* $i = 1, 2, 3$, *et* $m = 3$, *ainsi :*

$$H = \frac{12}{12\,(13)} \left( \frac{22^2}{4} + \frac{27^2}{4} + \frac{29^2}{4} \right) - 3\,(13) = 0.5.$$

*Au niveau de signification* $\alpha = 5\%$, *on a* $H = 0.5 \leq \chi^2_{(m-1)} = \chi^2_2 = 5.91$, *donc on accepte* $H_0 : F_1(x) = F_2(x) = F_3(x)$ .

**Code R :**

valeurs=c(34, 36, 41, 43, 44, 37, 45, 33, 35, 39, 42, 46)

groupes=factor(c("g1","g1","g1","g1","g2","g2","g2","g2","g3","g3","g3","g3"))

kruskal.test(valeurs ~ groupes)

Les résultats obtenus sont :

**Kruskal-Wallis chi-squared = 0.5, df = 2, p-value = 0.7788.**

On a la p-valeur est supérieure à 0.05, donc on accepte $H_0$.

## 2.2 Tests basés sur les distributions

Dans cette partie nous considérons trois types de tests statistique qui consiste à vérifier si l'échantillon observé suit une loi de probabilité théorique donnée. Ceci est formulé comme suit :

$$H_0 : F = F_0 \text{ contre } H_1 : F \neq F_0.$$

### 2.2.1 Test de Pearson d'ajustement (khi-deux)

Soit un échantillon de loi $P$ à valeurs dans un ensemble $O \in R$ tell que :

$$O = \cup_{k=1}^{m} O_k,\ O_k \cap O_j = \varnothing \quad k = 1, ..., m.$$

On définit le nombre $X_i$ appartenant à $O_k$ comme suit :

$$N_k = \sum_{i=1}^{n} \mathbb{I}_{(X_i \in O_k)},$$

Et les probabilités pour lesquelles $p_k$ on a $p_k = P(X_i \in O_k)$. La statistique du khi-2 d'ajustement est définie par :

$$Q := \sum_{k=1}^{m} \frac{(N_k - np_{0k})^2}{np_{0k}},$$

où $p_{0k} = P(O_k)$ sous $H_0$, on accepte $H_0 : (P_X = P_0)$ si $Q \leq \chi^2_{(m-1)}$, où $\chi^2_{(m-1)}$ le quantile de la statistique de khi-deux à $(m-1)$ degrés de liberté.

**Remarque 2.2.1** *Si la loi donnée $P_0$ avec un paramètre $r$ à estimer, alors le test du khi-deux à $(m-r-1)$ degrés de liberté.*

**Exemple 2.2.1** *Un pronostiqueur a observé les résultats de* 144 *couraes de chevaux .Le tableau suivant donne le nombre de vainqueurs selon les 8 positions de départ :*

| *Position de départ* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *Nombre de vainqueurs* | 29 | 19 | 18 | 25 | 17 | 10 | 15 | 11 |

*Pour simplifier l'écriture, on note* $P_i = P(x = i)$ *, pour* $i = 1, ..., 8$,*donc :*

$$P_1 + P_2 + P_3 + P_4 + P_5 + P_6 + P_7 + P_8 = 1$$

*Car les concurrents ont les mêmes chances de gagner, donc* $P_1 = P_2 = ... = 1/8$. *Pour* $m = 8$, $n = 144$ ,$P_k = 1/8$ *et* $nP_k = 18$ *pour tous* $k = 1, 2, ..., m$. *On récapitule les calculs dans le tableau suivant :*

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $N_k$ | 29 | 19 | 18 | 25 | 17 | 10 | 15 | 11 |
| $N_k - nP_k$ | 11 | 1 | 0 | 7 | $-1$ | $-8$ | $-3$ | $-7$ |
| $(N_k - nP_k)^2$ | 121 | 1 | 0 | 49 | 1 | 64 | 9 | 49 |
| $(N_k - nP_k)^2 / nP_k$ | 6.72 | 0.06 | 0 | 2.72 | 0.06 | 3.56 | 0.5 | 2.72 |

*Alors, le test statistique du khi-deux est égal à :*

$$Q = 6.72 + 0.06 + 2.72 + 0.06 + 3.56 + 0.5 + 2.72 = 16.34$$

*Pour* $\alpha = 5\%$, *on a* $Q = 16.34 \geq \chi^2_{(m-1)} = \chi^2_7 = 14.06$, *donc la position de départ a un effet sur les résultats de victoire.*

**Code R :**

obs=c(29, 19, 18, 25, 17, 10, 15, 11)

R=chisq.test(obs)

Les résultats sont :

**X-squared = 16.333, df = 7, p-value = 0.02224.**

Comme la p-valeur est inférieure à 0.05, donc on rejette $H_0 : F_X = F_0$.

### 2.2.2 Test de Kolmogorov-Smirnov

La statistique de test de Kolmogorov-Smirnov est définie par :

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

où $F_n$ est la fonction de répartition empirique associée à un echatillon $X_1, ..., X_n$. La distribution limite de la statistique $D_n$ est définie par :

$$\lim_{n \to \infty} P(\sqrt{n} D_n \leq x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^k \, e^{-2k^2 x^2} =: K(x), \text{ pour } x \geq 0,$$

et zero sinon, où $K(x)$ est appelée la fonction de répartition de Kolmogorov. Si note par $(X_{1:n}, ..., X_{n:n})$ la statistique d'ordre associée à l'échantillon $X_1, ..., X_n$ alors la statistique de Kolmogorov peut être rééctrite comme suit :

$$D_n := \max_{1 \leq i \leq n} \max \left\{ \left| F_0(X_{i:n}) - \frac{i}{n} \right|, \left| F_0(X_{i:n}) - \frac{i-1}{n} \right| \right\}.$$

On accepte $H_0 : F_X = F_0$ si $D_n \leq K^{-1}(1-\alpha)$ où $K^{-1}(1-\alpha)$ est le quantile d'ordre $(1-\alpha)$ de la statistique de Kolmogorov.

**Exemple 2.2.2** *Dans une étude de vibration, un échantillon aléatoire de* 13 *coposants d'avion a été soumis à de fortes vibration jusqu'à ce qu'ils présentent des*

*défaillances structrurelles. Les données fournies sont des temps de défaillance :*

$$17 \quad 11.4 \quad 4.6 \quad 19.5 \quad 8.8 \quad 22 \quad 11.6 \quad 9.5 \quad 5.7 \quad 13 \quad 17.3 \quad 15.7 \quad 6.3 \tag{2.1}$$

*On teste l'hypothèse nulle selon laquelle la fonction de distribution est :*

$$F(x) = \frac{x-4}{21}, \text{ pour } x \in [4, 25].$$

*Pour* $n = 13$*, on calcule la statistique de Kolmogorov-Smirnov comme suit :*

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 4.6 | 5.7 | 6.3 | 8.8 | 9.5 | 11.4 | 11.6 | 13 | 15.7 | 17 | 17.3 | 19.5 | 22 |
| $F(X_i)$ | 0.02 | 0.08 | 0.1 | 0.22 | 0.21 | 0.35 | 0.37 | 0.42 | 0.56 | 0.61 | 0.63 | 0.7 | 0.8 |
| $i/n$ | 0.07 | 0.15 | 0.23 | 0.30 | 0.38 | 0.46 | 0.53 | 0.61 | 0.69 | 0.76 | 0.84 | 0.92 | 1 |
| $i-1/n$ | 0 | 0.07 | 0.15 | 0.23 | 0.30 | 0.38 | 0.46 | 0.53 | 0.61 | 0.69 | 0.76 | 0.84 | 0.92 |
| $A$ | 0.05 | 0.1 | 0.13 | 0.08 | 0.17 | 0.11 | 0.16 | 0.21 | 0.13 | 0.15 | 0.21 | 0.22 | 0.2 |
| $B$ | 0.02 | 0.01 | 0.05 | 0.01 | 0.09 | 0.03 | 0.09 | 0.11 | 0.05 | 0.08 | 0.13 | 0.14 | 0.12 |

*où* $A = |F(X_i) - i/n|$ *et* $B = |F(X_i) - (i-1)/n|$*. Ainsi* $D_n = 0.21$*, D'après le tableau associé a la statisitque de Kolmogorov, le quantile, d'ordre* $(1-\alpha) = 0.95$*, noté* $K^{-1}(0.95) = 0.325$*. Pour* $n = 13$*, on a :* $D_n = 0.22 \leq 0.325$*, donc les données suivent une loi uniforme sur* $[4, 25]$*.*

**Code R :**

```
F=function (x){
ifelse(x<4,0,ifelse(x>25,1,(x - 4) / 21))
}
X=c(17,11.4,4.6,19.5,8.8,22,11.6,9.5,5.7,13,17.3,15.7,6.3)
```

ks.test(X,punif,max=25,min=4)

Donc les résultats donné par :

**D = 0.21, p-value = 0.5294.**

Comme la p-valeur est supérieure à 0.05, donc on accepte $H_0 : F_X = F_Y$.

### 2.2.3 Test de Lilliefors

Ce test étudie l'ajustement de normalité c'est à dire :

$$H_0 : F(x) = \Phi_{\mu,\sigma^2}(x),$$

où $\Phi_{\mu,\sigma^2}(\cdot)$ est la fonction de répartition de la loi normale. On a définit la statistique du test de Lilliefors comme suit :

$$L_n := \max_{1\leq i\leq n} \max\left\{\left|F_0(\frac{X_{i:n}-\overline{X}}{S}) - \frac{i}{n}\right|, \left|F_0(\frac{X_{i:n}-\overline{X}}{S}) - \frac{i-1}{n}\right|\right\}.$$

Avec :

▷ $\overline{X}$ : Est la moyenne empirique de $X_{i:n}$.

▷ $S$ : Est l'écart-type empirique de $X_{i:n}$.

On accepte l'hypothèse $H_0$ si $L_n \leq l_\alpha$ ,où $l_\alpha$ est le quantile d'ordre $\alpha$ de Lilliefors.

**Exemple 2.2.3** *On choisit l'exemple dans (2.1) mais ici, $F(x)$ est une loi normale centrée réduite, avec une taille $n = 11$. Nous avos resumé les résultats dans*

*le tableau suivant :*

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $A$ | $-1.49$ | $-1.28$ | $-1.17$ | $-0.69$ | $-0.56$ | $-0.2$ | $-0.16$ | 0.09 | 0.6 | 0.85 | 0.9 |
| $B$ | 0.06 | 0.10 | 0.12 | 0.24 | 0.29 | 0.42 | 0.43 | 0.53 | 0.72 | 0.8 | 0.81 |
| $C$ | 0.01 | 0.05 | 0.11 | 0.06 | 0.09 | 0.04 | 0.1 | 0.08 | 0.03 | 0.03 | 0.03 |
| $D$ | 0.06 | 0.03 | 0.03 | 0.01 | 0.01 | 0.04 | 0.03 | 0 | 0.11 | 0.01 | 0.05 |

*où*

$$A = \frac{X_{i:n} - \overline{X}}{S},\ D = \left| F\left(\frac{X_{i:n} - \overline{X}}{S}\right) - \frac{i-1}{n} \right|,$$
$$B = F\left(\frac{X_{i:n} - \overline{X}}{S}\right)\ \text{et}\ C = \left| F\left(\frac{X_{i:n} - \overline{X}}{S}\right) - \frac{i}{n} \right|,$$

*avec* $\overline{X} = 12.49$ *et* $S = \sqrt{28.07} = 5.29$. *Nous avons calculé cette valeur de la manière suivante : pour* $i = 1$ *on a*

$$F\left(\frac{X_1 - \overline{X}}{S}\right) = F\left(-1.49\right) = 0.06,$$

*d'après la table de la loi normale centrée réduite (côté négatif). De la même manière, nous procédons pour* $i = 1, ..., 11$, *en répétant cette méthode pour les autres valeurs. Ainsi* $L_n = 0.11$.

*Pour* $\alpha = 5\%$, *ona* $L_n = 0.12 \leq l_\alpha = 0.24$, *les données suivent une loi normale centrée réduite.*

**Code R :**

```
library(nortest)
x=c(4.6, 5.7, 6.3, 8.8, 9.5, 11.4, 11.6, 13, 15.7, 17, 17.3)
```

lillie.test(x)

Nous obtenons ce qui suit :

**D = 0.126, p-value = 0.894**

Donc, la p-valeur est supérieure à 0.05, donc on accepte $H_0 : F_X = \Phi_{0,1}(x)$

### 2.2.4 Test de Cramer-von Mises d'ajustement

On définit cette statistique sous $H_0 : F_X = F_0$ comme suit :

$$W_n := n \int_{\mathbb{R}} (F_n(x) - F(x))^2 dF(x) = \sum_{i=1}^{n} F_0(X_{i:n} - \frac{2i-1}{2n})^2 + \frac{1}{12n}.$$

On accepte l'hypothèse $H_0$ si $W_n \leq w_\alpha$, où $w_\alpha$ est le quantile du test de Cramer-von Mises d'ordre $\alpha$.

**Exemple 2.2.4** *Supposons que ce sont les temps d'attente (en minutes) pour un certain nombre de clients dans la file d'attente du supermarché.*

3.2, 4.5, 2.8, 5.1, 3.9, 4.0, 2.5, 3.6, 4.2, 5.0, 3.7, 2.9, 4.3, 3.1, 5.3

*Tester si un petit échantillon suit une loi exponentielle de paramètre $\lambda = 2$.*

**Code R :**

library(goftest)

x=c(3.2, 4.5, 2.8, 5.1, 3.9, 4.0, 2.5, 3.6, 4.2, 5.0, 3.7, 2.9, 4.3, 3.1, 5.3)

cvm.test(x, null = "pexp", rate = 2)

Les résultats obtenus sont :

**omega2 = 4.407, p-value < 2.2e-16.**

Comme la p-valeur est inférieure à 0.05.Donc, les données ne suivent donc pas une loi exponentielle.

### 2.2.5 Test de Shapiro-Wilk

Ce test permet d'étudier si $F_X$ suit une distribution normale de paramètres $(\mu, \delta^2)$ .
La statistique correspondante est définie comme suit :

$$W := \frac{\sum_{i=1}^{n} a_i X_{i:n}}{\sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

Où :

- $a_i = (a_1, a_2, ..., a_n) = \left(m^T V^{-1}\right) \left(m^T V^{-1} V^{-1} m\right)^{-1/2}$ .
- $m_i = (m_1, m_2, ..., m_n)$ est l'espérance de la statistique d'ordre $(X_{1:n}, X_{2:n}, ..., X_{n:n})$ et $V$ est la matrice de covariance de cette statistique d'ordre.

**Exemple 2.2.5** *Dans le logiciel R, on a :*

n=20

y=rnorm(n,mean=20,sd=2)

shapiro.test(y)

Aprés l'excusions,voici les résultas obtenus :

**W = 0.96042, p-value = 0.5523.**

Donc, la p-valeur est supérieure à 0.05, donc les données suivent une distribution normale $N(\mu = 20, \delta^2 = 4)$ .

### 2.2.6 Test d'Anderson-Darling d'ajustement

Le fondement du test Anderson-Darling porte sur l'hypothèse $H_0 : F = F_0$. La statistique correspondante est définie comme suit :

$$A := -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\ln F_0\left(X_{i:n}\right) + (2n+1-2i)\ln\left(1-F_0\left(X_{i:n}\right)\right)\right].$$

On accepte $H_0$ si $A \leq a_\alpha$, où $a_\alpha$ est le quantile d'ordre $\alpha$ de test Anderson-Darling.

**Exemple 2.2.6** *Soit l'échantillon suivant de taille* $n = 8$ :

$$7.8 \quad 9 \quad 9.7 \quad 8.3 \quad 5.7 \quad 6.4 \quad 7.2 \quad 11$$

*Il s'agit de tester si* $F = \Phi_{0,1}$ *(gaussienne centrée-réduite). Pour* $n = 8$, *nous avons* $\overline{X} = 8.15$ *et* $S = \sqrt{3.07} = 1.75$. *On note par* $Y_{i:n} := \left(X_{i:n} - \overline{X}\right)/S$, $N_1 = 2i-1$, $N_2 = 2n+1-2i$, $a = \ln\left(\Phi_{0,1}\left(Y_{i:n}\right)\right)$, $b = \ln\left(1-\Phi_{0,1}\left(Y_{i:n}\right)\right)$ *et*

$$T_i := (2i-1)\ln\Phi_{0,1}\left(Y_{i:n}\right) + (2n+1-2i)\ln\left(1-\Phi_{0,1}\left(Y_{i:n}\right)\right).$$

*Le tableau suivant récapitule les valeurs de* $Y_{i:n}$ *et* $T_i$ :

| $i$ | $X_{i:n}$ | $Y_{i:n}$ | $N_1$ | $N_2$ | $\Phi_{0,1}(Y_{i:n})$ | $a$ | $1-\Phi_{0,1}(Y_{i:n})$ | $b$ | $T_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5.7 | $-1.4$ | 1 | 15 | 0.08 | $-2.51$ | 0.91 | $-0.08$ | $-3.71$ |
| 2 | 6.4 | $-1$ | 3 | 13 | 0.15 | $-1.84$ | 0.84 | $-0.17$ | $-7.77$ |
| 3 | 7.2 | $-0.54$ | 5 | 11 | 0.29 | $-1.22$ | 0.7 | $-0.34$ | $-9.96$ |
| 4 | 7.8 | $-0.2$ | 7 | 9 | 0.42 | $-0.86$ | 0.57 | $-0.54$ | $-10.90$ |
| 5 | 8.3 | 0.08 | 9 | 7 | 0.53 | $-0.62$ | 0.46 | $-0.76$ | $-10.95$ |
| 6 | 9 | 0.48 | 11 | 5 | 0.68 | $-0.37$ | 0.31 | $-1.16$ | $-9.95$ |
| 7 | 9.7 | 0.88 | 13 | 3 | 0.81 | $-0.2$ | 0.18 | $-1.67$ | $-7.8$ |
| 8 | 11 | 1.62 | 15 | 1 | 0.94 | $-0.05$ | 0.05 | $-2.95$ | $-3.77$ |

*Ainsi la valeur de la statisitique d'Aderson-Darling égale à :*

$$A = -8 - \frac{1}{8}(-3.71 - 7.77 - 9.96 - 10.90 - 10.95 - 9.95 - 7.8 - 3.77) = 0.10.$$

*Alors pour* $\alpha = 5\%$, *on a* $A = 0.10 < a_\alpha = 0.787$, *donc les données suivent une loi gaussienne centrée-réduite.*

**Code R :**

```
library(nortest)
x=c(5.7, 6.4, 7.2, 7.8, 8.3, 9, 9.7, 11)
ad.test(x)
```

Les résultats donnés par :

**A = 0.10, p-value = 0.9873.**

Pour la p-valeur supérieure à $0,05$, on accepte $H_0 : F_X = F_0$.

### 2.2.7 Test de Pearson (khi-deux) d'homoegéneité

Le test de Test de Pearson est utilisé pour comparer si deux échantillons (indépendants) $X := (X_1, ..., X_{n_1})$ et $Y := (Y_1, ..., Y_{n_2})$ suivent la même distribution. Supposons que l'échantillon $(X; Y)$ prend ces valeurs dans l'ensemble $\{a_1, a_2, ..., a_m\}$, où $1 \leq m \leq n$ avec $n := n_1 + n_2$. On définit le test du Khi-deux d'homogeneité comme suit :

$$Q_{\text{hom}} := \sum_{k=1}^{m} \left[ \frac{n_1 \left( \frac{N_k M_k}{n} - \frac{N_k}{n_1} \right)^2}{\frac{N_k M_k}{n}} + \frac{n_2 \left( \frac{N_k M_k}{n} - \frac{M_k}{n_2} \right)^2}{\frac{N_k M_k}{n}} \right],$$

où **:**

- $N_k$ : est le nombre d'observations de $(X_1, ..., X_{n_1})$ prennenant la valeur $a_k$.
- $M_k$ : est le nombre d'observations de $(Y_1, ..., Y_{n_2})$ qui prennent la valeur $a_k$.

On accepte l'hypothèse $H_0$ si $Q_{\text{hom}} \leq \chi^2_{(m-1)}$, où $\chi^2_{(m-1)}$ est le quantile de la loi du khi-deux à $(m-1)$ degrés de liberté .

**Exemple 2.2.7** *Il s'agit de tester si la préférence de Boisson (Café,Thé, Jus) est la même chez les hommes et les femmes :*

| *Boisson* | *Café* | *Thé* | *Jus* |
|---|---|---|---|
| *Hommes* | 30 | 10 | 10 |
| *Femmes* | 20 | 25 | 15 |

*Dans cet example, nous avons :*

$$m = 3,\ n_1 = 50,\ n_2 = 60,\ n = \ n_1 + n_2 = 110,$$

$$N_1 = 30,\ N_2 = 10,\ N_3 = 10,\ M_1 = 20,\ M_2 = 25\ et\ M_3 = 15.$$

*Donc, on calcule la statistique du khi-deux comme suit*

$$\begin{aligned} Q_{\text{hom}} &= \frac{50\left((30\times 20)/110 - 30/50\right)^2}{(30\times 20)/110} + \frac{60\left((30\times 20)/110 - 20/60\right)^2}{(30\times 20)/110} \\ &+ \frac{50\left((10\times 25)/110 - 10/50\right)^2}{(10\times 25)/110} + \frac{60\left((10\times 25)/110 - 25/60\right)^2}{(10\times 25)/110} \\ &+ \frac{50\left((10\times 15)/110 - 10/50\right)^2}{(10\times 15)/110} + \frac{60\left((10\times 15)/110 - 15/60\right)^2}{(10\times 15)/110} \\ &= 794.20. \end{aligned}$$

*Pour* $\alpha = 5\%$,*on a* $Q_{\text{hom}} = 794.20 \geq \chi^2_2 = 0.10$, *alors on rejette* $H_0$. *En d'autres termes, il y a une différence significative dans la répartition des préférences de boisson entre hommes et femmes.*

**CodeR :**

tabl=matrix(c(30, 10, 10, 20, 25, 15),nrow=2,byrow=TRUE)

rownames(tabl)= c("homme","femme")

colnames(tabl)= c("Café","Thé","Jus")

chisq.test(tabl)

Les résultat sont :

**X-squared = 8.5905, df = 2, p-value = 0.01363.**

Comme la p-valeur est inférieure à 0.05, donc on rejette $H_0 : F_X = F_Y$.

### 2.2.8 Test de Cramer-von Mises d'homogéneité

Le test de Cramér–von Mises est aussi utilisé pour comparer si deux échantillons (indépendants) $X := (X_1, ..., X_n)$ et $Y := (Y_1, ..., Y_m)$ suivent la même distribution. Sa statistique est définie comme suit :

$$T := \frac{nm}{(n+m)^2} \sum_{i=1}^{n+m} (F_n(Z_i) - G_m(Z_i))^2,$$

où $Z := \{X_1, ..., X_n; Y_1, ..., Y_m\} =: \{Z_1, ..., Z_k\}$, où $k := n + m$, et $F_n$ et $G_m$ sont les fonctions de répartition empirique associées aux deux échantillons $X$ et $Y$ respectivement. En termes de rangs statistiques cette formule peut être réécrite comme suit :

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

où

$$U := n \sum_{i=1}^{n} (r_i - i)^2 + m \sum_{j=1}^{m} (s_j - j)^2,$$

avec $r_i$ et $s_j$ designent, respectivement, les rangs des statistiques d'ordre $X_{i:n}$ et $Y_{j:m}$ dans l'échantillon ordonné $Z_{1:k} \leq ... \leq Z_{k:k}$. On accepte $H_0$ si $T \leq w_\alpha$, où $w_\alpha$ est le quantile du test de Cramer–von Mises d'ordre $\alpha$.

**Exemple 2.2.8** *On s'interesse à tester si les deux échantillons suivants proviennent de la même distribution :*

| | | | | | | |
|---|---|---|---|---|---|---|
| $X$ | $-0.26$ | $0.42$ | $-0.52$ | $0.56$ | $-1.11$ | $-0.40$ |
| $Y$ | $-1.38$ | $1.22$ | $0.64$ | $0.28$ | $-0.75$ | / |

Pour cet exemple $n = 6$, $m = 5$. Les valeurs de $Z_i$, $F_n(Z_i)$, $G_m(Z_i)$ et $d_i^2 := (F_n(Z_i) - G_m(Z_i))^2$ sont récapitulées au tableau suivant :

| $Z_i$ | $-0.26$ | $0.42$ | $-0.52$ | $0.56$ | $-1.11$ | $-0.40$ | $-1.38$ | $1.22$ | $0.64$ | $0.28$ | $-0.75$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $6F_n(Z_i)$ | 4 | 5 | 2 | 6 | 1 | 3 | 0 | 6 | 6 | 4 | 1 |
| $5G_m(Z_i)$ | 2 | 3 | 2 | 3 | 1 | 2 | 1 | 5 | 4 | 3 | 2 |
| $d_i$ | $\frac{4}{15}$ | $\frac{7}{30}$ | $-\frac{1}{15}$ | $\frac{2}{5}$ | $-\frac{1}{30}$ | $\frac{1}{10}$ | $-\frac{1}{5}$ | 0 | $\frac{1}{5}$ | $\frac{1}{15}$ | $-\frac{7}{30}$ |

Un calcul élémentaire donne $\sum_{i=1}^{11} d_i^2 = 0.44$, et par censéquent :

$$T = \frac{5 \times 6}{(5+6)^2} 0.44 \simeq 0.1091.$$

Pour $\alpha = 5\%$, on a $W_n = 0.1091 < w_\alpha \simeq 0.468$, alors, les deux échantillons suivants proviennent de la même distribution.

En termes de rangs :

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_{i:n}$ | $-1.11$ | $-0.52$ | $-0.40$ | $-0.26$ | $0.42$ | $0.56$ | $*$ | $*$ | $*$ | $*$ | $*$ |
| $Y_{i:m}$ | $-1.38$ | $-0.75$ | $0.28$ | $0.64$ | $1.22$ | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ |
| $Z_{i:k}$ | $-1.38$ | $-1.11$ | $-0.75$ | $-0.52$ | $-0.40$ | $-0.26$ | $0.28$ | $0.42$ | $0.56$ | $0.64$ | $1.22$ |
| $r_i$ | 2 | 4 | 5 | 6 | 8 | 9 | $*$ | $*$ | $*$ | $*$ | $*$ |
| $s_j$ | 1 | 3 | 7 | 10 | 11 | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ |
| $r_i - i$ | 1 | 2 | 2 | 2 | 3 | 3 | $*$ | $*$ | $*$ | $*$ | $*$ |
| $s_j - j$ | 0 | 1 | 4 | 6 | 6 | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ |
| $(r_i - i)^2$ | 1 | 4 | 4 | 4 | 9 | 9 | $*$ | $*$ | $*$ | $*$ | $*$ |
| $(s_j - j)^2$ | 0 | 1 | 16 | 36 | 36 | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ |

Ceci donne :

$$\sum_{i=}^{6} (r_i - i)^2 = 31 \text{ et } \sum_{j=1}^{5} (s_j - j)^2 = 89.$$

Par conséquent $U = 6 \times 31 + 5 \times 89 = 513 = 631$ et finalement

$$T = \frac{631}{6 \times 5 \times (6+5)} - \frac{4 \times 6 \times 5 - 1}{6\,(6+5)} \simeq 0.1091.$$

**Code R :**

library(cramer)

x = c(-0.26, 0.42, -0.52, 0.56, -1.11, -0.4)

y = c(-1.38, 1.22, 0.64, 0.28, -0.75)

cramer.test(x, y)

Les résultats sont :

**Cramer-von Mises Test**

**data : x and y**

**W = 0.107, p-value = 0.562.**

Pour p-valeur est supérieure à 0.05, donc on accepte $H_0 : F_X = F_Y$.

### 2.2.9 Test d'indépendance du khi-deux

Soient $X$ et $Y$ deux variables aléatoires et suppose qu'on dispose de $n$ observations de $(X, Y)$. Divisons l'espace des valeurs de $X$ (la droite réelle) en $r$ intervalles disjoints $A_1, ..., A_r$. De même on divise l'espace des valeurs de $Y$ (la droite réelle) en $c$ intervalles disjoints $B_1, ..., B_c$. En règle générale, on choisit la longueur de chaque intervalle de telle sorte que la probabilité que $(X, Y)$ se situe dans un intervalle soit approximativement $(1/r)\,(1/c)$. En outre il est important que $n/r$

et $n/c$ soient au moins égaux à 5. Soit $X_{ij}$ le nombre de paires $(X_k, Y_k)$, $k = 1, ..., n$, qui tombent dans $A_i \times B_j$ et soit

$$p_{ij} := P\left((X, Y) \in A_i \times B_j\right) = P\left(X \in A_i \text{ et } Y \in B_j\right),$$

où $i = 1, ..., r$ et $j = 1, ..., c$. Si chaque $p_{ij}$ est déterminé, la quantité

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(X_{ij} - np_{ij}\right)^2}{np_{ij}}$$

suit approximativement la loi du khi-deux $rc - 1$ degrès de liberté, à condition que $n$ soit grand. Si $X$ et $Y$ sont indépendantes alors $P\left((X, Y) \in A_i \times B_j\right) = P\left(X \in A_i\right) P\left(Y \in B_j\right)$. On pose $p_{i\cdot} := P\left(X \in A_i\right)$ et $p_{\cdot j} := P\left(Y \in B_j\right)$. Sous l'hypothèse nulle $H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}$, $i = 1, ..., r$ et $j = 1, ..., c$. En pratique $p_{ij}$ n'est pas connue alors on remplace cette dernière par son estimateur approprié. Sous $H_0$ on estime $p_{i\cdot}$ par

$$\widehat{p}_{i\cdot} := \frac{\sum_{i=1}^{r} X_{ij}}{n},\ i = 1, ..., r$$

et $p_{\cdot j}$ par

$$\widehat{p}_{\cdot j} := \frac{\sum_{j=1}^{c} X_{ij}}{n},\ j = 1, ..., c.$$

Puisque $\sum_{i=1}^{r} \widehat{p}_{i\cdot} = 1 = \sum_{j=1}^{c} \widehat{p}_{\cdot j}$, nous avons estimer que $r - 1 + c - 1 = r + c - 2$ paramètres. Par conséquent, sous l'hypothèse nulle $H_0$, la variable aléaroire :

$$Q_{ind} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(X_{ij} - np_{ij}\right)^2}{n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}}$$

suit, asymptotiquement, la loi du khi-deux à $rc - 1 - (r + c - 2) = (r - 1)(c - 1)$ degrés de liberté. L'hypothèse null $H_0$ est rejetée si la valeur de $Q_{ind} > \chi^2_{(r-1)(c-1),\alpha}$, où $\chi^2_{(r-1)(c-1),\alpha}$ est la quantile d'ordre $\alpha$ de loi du khi-deux à $(r - 1)(c - 1)$ degrés

de liberté.

**Exemple 2.2.9** *Nous voulons connaître l'effet du sport sur le nombre de personnes en bonne ou en mauvaise santé, et nous prenons l'exemple suivant :*

| | *En bonne santé* | *Au mauvaise santé* | |
|---|---|---|---|
| *Pratique le sport* | 40 | 15 | $n\widehat{p}_{1*} = 55$ |
| *Ne pratique pas le sport* | 20 | 25 | $n\widehat{p}_{2*} = 45$ |
| | $n\widehat{p}_{*1} = 60$ | $n\widehat{p}_{*2} = 40$ | |

*Nous avons* $r = 2, c = 2,$ *et* $n = 100,$ *ainsi :*

$$Q_{ind} = \frac{\left(\frac{55\times60}{100} - 40\right)^2}{(55 \times 60)/100} + \frac{\left(\frac{55\times40}{100} - 15\right)^2}{(55 \times 40)/100} + \frac{\left(\frac{45\times60}{100} - 20\right)^2}{(45 \times 60)/100} + \frac{\left(\frac{45\times40}{100} - 25\right)^2}{(45 \times 40)/100} = 7.12.$$

*Au niveau de signification* $\alpha = 5\%,$ *on a* $Q_{ind} = 7.12 \geq \chi^2_1 = 3.84,$ *donc il existe une relation entre la pratique du sport et l'état de santé.*

**Code R :**

tabl=matrix(c(40,15,20,25),nrow=2,byrow=TRUE)

dimnames = list(c("Pratique le sport","Ne pratique pas le sport"),c("En bonne santé","Au mauvaise santé"))

chisq.test(tabl)

Donc, on a :

**X-squared = 7.118, df = 1, p-value = 0.007654.**

Pour la p-valeur est inférieure à 0.05, donc on rejette $H_0$ : $X$ est indépendant de $Y$.

## 2.3 Comparaison entre les tests non paramétriques

Après avoir présenté une série de tests non paramétriques à travers des exemples pratiques précédents, il est désormais possible d'effectuer une comparaison graphique illustrant les différences de performance entre ces tests. La figure suivante montre une comparaison de leur puissance statistique, réalisée au niveau de signification $\alpha = 0.05$.

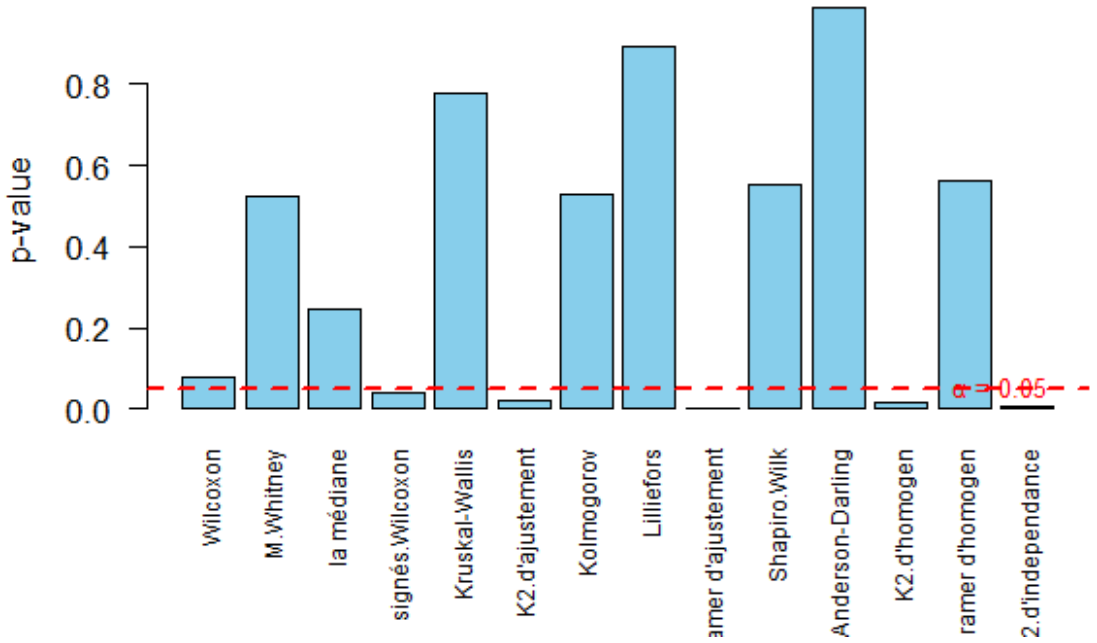


FIG. 2.1 – Comparaison graphique entre les tests non paramétriques présentés précédemment en termes de puissance statistique pour $\alpha = 0.05$

Cette comparaison met en évidence la capacité de chaque test à détecter des différences réelles entre les groupes étudiés, ce qui permet de choisir l'outil statistique le plus approprié selon la nature des données et le contexte de l'analyse .

# Conclusion

L'ajustement d'un modèle sur un échantillon consiste à déterminer si le modèle statistique choisi s'adapte bien aux données observées dans l'échantillon. Pour effectuer ceci nous avons deux méthodes : l'ajustement graphique et l'ajustement par les tests non-paramétriques. La première méthode est basée sur les estimateurs non-paramétriques de la densité de probabilité et quantiles afin de comparer ceux-ci avec la densité ou la fonction des quantiles à ajuster. La deuxième méthode utilise les tests non-paramétriques classiques à savoir le test de Kolmogorov-Smirnov, test de Cramer-Von Mises, test de Khi-deux,....

# Bibliographie


[1] Al-Kenani, A. , & Yu, K.(2012). New Bandwidth Selection for Kernel Quantile Estimators.Journal of Probability and Statistics.

[2] Anderson, T. W. (2010). Anderson-Darling tests of goodness of fit. Stanford University.

[3] Berlinet, A. & Devroye, L. (1989). Estimation d'une densité : un point sur la méthode du noyau. Statistique et analyse des données, $14(1), 1 - 32$.

[4] Bhar, L.Nonparametric tests. Indian Agricultural Statistics Research Institute. Library Avenue, New Delhi-12. lmb@iasri.res.in.

[5] Brown, GW ; Mood, AM (June1948). "Homogeneity of several samples". American Statistician.$2(3) : 22 - 23$.

[6] Colletaz, G. (2021, March 22). Statistique non paramétrique. Master 2 Économétrie et Statistique Appliquée.

[7] Corder, G.W. & Foreman, D.I. (2014). Nonparametric Statistics : A Step-by-Step Approach, Wiley. ISBN 978-1118840313.

[8] DATAtab. (2021, 3 may). Kruskal-Wallis-Test (simply explained) [Video]. YouTube.https ://www.youtube.com/watch ?v=l86wEhUzkY4.

[9] Deheuvels, P.(1977) : Estimation non paramétrique de la densité par histogrammes généralisés. Revue de statistique appliquée, $25(3), 5 - 42$.

[10] David, H. A., & Nagaraja, H. N. (2003). Order Statistics (2nd ed.). Wiley-Interscience.

[11] Falk, M. (1984). Relative Deficiency of Kernel Type Estimators of Quantiles, The Annals of Statistics.

[12] Gonzalez, T. F., Sahni, S., & Franta, W. R.(1977). An efficient algorithm for the Kolmogorov–Smirnov and Lilliefors tests. ACM Transactions on Mathematical Software, $3(1), 60 - 64$.

[13] Henkouche, M. Statistique non paramétrique. Département de mathématiques,U.S.T.O.M.B.

[14] Lopes, R. H. C., Reid, I. D., & Hobson, P. R. (2007). The two-dimensional Kolmogorov–Smirnov test. Proceedings of Science, ACAT 2007(045), $1 - 12$.

[15] Matias, C.(2012, *September*). Introduction à la statistique non paramétrique. Atelier SFDS, CNRS, Laboratoire Statistique & Génome, Évry. Retrieved from $27 - 28$ .

[16] Marron, J.S. & Ruppert, D. (1994). Transformations to reduce boundarybias in kernel density estimation. Journal of the Royal Statistical Society $653-671$.

[17] Mood, A.M., (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. Ann. Math. Statist. $514 - 522$.

[18] Parzen, E. (1962). On estimation of a probability density function and mode.Annals of Mathematical Statistics $33, 1065 - 1076$.

[19] Rakotomala, R.,(2008). Comparaison de populations. Tests Non Paramétriques.

[20] Razali, N. M., & Yap, B. W. (2011).Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, $2(1), 21 - 33$.

[21] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27(3), 832–837.

[22] Ruch, J.-J. (2013). Statistique :Tests d'hypothèses. Préparation à l'Agrégation Bordeaux 1.

[23] Sheather, S. J., & Marron, J. S. (1990). Kernel quantile estimators. Journal of the American Statistical Association,85(410), $410 - 416$..

[24] Vinatier, S. (2008). Compléments de mathématiques. Licence de Biologie, $3^e$ semestre, Faculté des Sciences et Techniques de Limoges.

[25] Wikipédia.Test de Kruskal-Walis , https ://fr.wikipedia.org/wiki/Test-de-Kruskal-Wallis

## ملخص

تكلمنا في هذه المذكرة عن موضوع ضبط نماذج الاحتمال حيث قمنا بدراسة الطرق اللامعلمية بشكل خاص وركزنا على طريقتنين الاولى تقدير الكثافة بالمدرج التكراري وطريقة النواة,الى جانب دراسة الكونتيلات كما تطرقنا الى الطريقة الثانية وهي الاختبارات اللامعلمية المبنية على الرتب والتوزيعات .اما من الناحية التطبيقية فقد استخدمنا محاكاة البيانات لمقارنة النتائج النظرية بالتقديرات التجريبية.



## Résumé

Nous avons abordé dans ce mémoire le thème de l'ajustement des modèles de probabilité, en nous concentrant particulièrement sur les méthodes non paramétriques. Nous avons mis l'accent sur deux approches principales : l'estimation de la densité à l'aide de l'histogramme et du noyau, ainsi que l'étude des quantiles. Nous avons également traité une deuxième méthode, à savoir les tests non paramétriques basés sur les rangs et les distributions. Sur le plan pratique, nous avons utilisé la simulation de données afin de comparer les résultats théoriques aux estimations empiriques.



## Abstract

In this thesis, we addressed the topic of probability model fitting, with a particular focus on non-parametric methods. We emphasized two main approaches: density estimation using histograms and kernel methods, as well as the study of quantiles. We also discussed a second approach involving non-parametric tests based on ranks and distributions. From a practical standpoint, we used data simulation to compare theoretical results with empirical estimates.