République Algériennne 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

Départementde Mathématiques.

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathematiques Appliquées**"

Option : Statistiques .

Par Mr. DHIAB Mohammed Zakaria

Titre :

Tests d'ajustement et applications

Devant le Jury :

Mme.	TOUBA Sonia	MCB	U. Biskra	Président
Mlle.	ROUBI Affef	MAA	U. Biskra	Encadreur
Mme.	BENBRIKA Ghozlane	MAA	U. Biskra	Examinatrice

Soutenu Publiquement le 28/06/2022

Dédicace

Ce travail est dédié à

A mes Parents pour tout le soutien qu'ils m'ont apportés, toujours été à mes côtés pour m'encourager. Que ce travail traduit ma gratitude et mon affection.

A mes chers grands-parents, Qui je leur souhaite une bonne santé.

A mes sœurs pour leur présence à mes côtés.

A tous mes amis qui m'ont encouragé dans les moments difficiles.

A tous ceux qui de près ou de loin m'ont soutenu.

A tous ceux que j'aime et ceux qui m'aiment.

Remerciements

C'est avec un grand plaisir que je remercie Dieu tout puissant qui m'a donné la force et la santé pour que ce mémoire soit mené à son terme.

Je tiens aussi à remercier **Dr ROUBI Affef** mon encadreur de mémoire et professeur à l'université Université Mohamed Khider pour m'avoir accordé toute sa confiance mais aussi pour tout le soutient et les conseils avisés pour l'avancement de la mémoire.

Je remercie les membres du jury qui me font honneur en jugeant ce travail.

A ma famille et mes proches, notamment pour leur soutien moral, je vous suis très reconnaissant.

A Tous ceux et celles qui ont participé de près ou de loin à l'élaboration de ce travail, qu'ils trouvent ici l'expression de ma haute considération

Table des matières

Liste des tableaux

Introduction

Le problème qu'on va l'examiner dans ce mémoire est d'une grande importance pratique : Etant donné un échantillon observé $x_1, x_2, ..., x_n$ constitué d'une suite numérique de mesures indépendantes d'un phénomène aléatoire dont la loi de probabilité n'est pas connue précisément, on veut tester si cet échantillon provient d'une loi F donnée par exemple de la loi $N(0,1)$. Les méthodes qu'on va utiliser s'appellent méthodes d'ajustement de l'échantillon observé à la loi théorique F.

Ce mémoire est composé de deux chapitres.

Chapitre 1 : dans ce chapitre, on donné quelques notions générales sur les tests statistique tels que hypothèse, risque, région critique, ..., aussi on a présenté les différents propiriétés d'un test statistique et ses catégories.

Chapitre 2 : ce chapitre est consacré aux tests statistique d'ajustement, dont on a présenté parmi ceux le plus courants : comme le test du khi-deux, de Kolmogorov-Smirnov, ... , aussi on a donné quelques exemples d'application de ces derniers.

Chapitre 1

Notions générales sur les tests statistiques

Dans cette section, on va donner tous les éléments d'un test d'hypothése.

Soit le modèle $(\Omega, \Lambda, P_0, \theta \in \Theta)$, on définit $\Theta = \theta_0 \cup \theta_1$ et les hypothèses H_0 par $\theta \in \theta_0$ et H_1 par $\theta \in \theta_1$ qui on va les tester contre eux pour décider sur la valeur de θ.

1.1 Hypothèses

1.1.1 Hypothèse nulle et L'hypothèse alternative

Définition 1.1.1 *Une **hypothèse nulle** notée H_0 est l'hypothèse que l'on désire contrôler : elle consiste par exemple à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et est due aux fluctuations d'échantillonnage. Cette hypothèse est formulée dans le but d'être rejetée [14].*

D'autre part ***L'hypothèse alternative*** *notée H_1 est la négation de H_0, elle est équivalente à dire «H_0 est fausse». La décision de rejeter H_0 signifie que H_1 est réalisée ou H_1 est vraie [14].*

1.1.2 Hypothèse composite

Définition 1.1.2 *On dite que H_j composite si Θ_j n'est pas réduit un singleton, sinon elle est simple.*

1.2 Test d'hypothèse

Définition 1.2.1 *Soit la fonction $\Phi(X_1, \ldots, X_n)$ à valeurs dans $\{0, 1\}$, Φ est mesurable et est calculée d'un n-échantillon i.i.d issu de la variable aléatoire X de loi P_0, cette fonction est appelée test de l'hypothèse H_0 contre H_1.*

On maintient H_0 si $\Phi(X) = 0$, si non, on maintient H_1.

1.3 Erreur et risque

1.3.1 Erreur

Définition 1.3.1 *La décision d'un test se base sur les données d'un échantillon aléatoire de la population. Il y a donc deux types d'erreurs possibles dans un test statistique*

1. L'erreur de type I qui consiste à rejeter H_0 alors que H_0 est vraie.

2. L'erreur de type II qui consiste à accepter H_0 alors que H_0 est fausse.

1.3.2 Risque

Définition 1.3.2 *Dans le contexte d'un test d'hypothèses, on appelle risque la probabilité de commettre une erreur. Puisqu'il y a deux types d'erreurs, on distingue donc deux types de risques qu' on les note α et β*

$$\begin{aligned}\alpha &= P\ (Erreur\ de\ type\ I) = P\ (rejeter\ H_0/H_0\ est\ vraie).\\ &= P\ (H_1/H_0).\end{aligned}$$

tq : α est risque de première espèce.

$$\begin{aligned}\beta &= P\ (Erreur\ de\ type\ II) = P\ (accepter\ H_0/H_0\ est\ fausse).\\ &= P\ (H_0/H_1).\end{aligned}$$

tq : β est risque de deuxième espèce.

1.3.3 Puissance

Définition 1.3.3 *[2]*

La puissance d'un test est la probabilité de rejeter l'hypothèse nulle H_0 quand l'alternative H_1 est vraie. On la note par

$$\pi := P[rejeter\ H_0|H_1\ est\ vraie] = 1-\beta.$$

Lorsque H_1 est composite, la puissance est variable sur Θ_1. De même lorsque H_0 est composite, le risque de première espèce est variable sur Θ_0. On définit alors

une fonction sur l'ensemble Θ *qu'on appelle fonction puissance*

$$\pi(\theta) := P_0[rejeter\ H_0],\ \theta \in \Theta.$$

Remarque 1.3.1 *1- Si* $\theta \in \Theta_0, \pi(\theta) = \alpha(\theta)$ *c'est le risque de première espèce.*

2- Si $\theta \in \Theta_1, \pi(\theta) = 1 - \beta(\theta)$ *c'est la puissance du test.*

1.3.4 Statistique du test

Définition 1.3.4 *La statistique qui apporte le plus de renseignement sur le problème posé est appelée variable de décision ou statistique du test. La loi de probabilité doit être différente selon que* H_0 *ou* H_1 *; sinon elle ne servait à rien [2].*

1.3.5 Région de rejet et région critique

Définition 1.3.5 *[2]*

La région de rejet d'un test est l'ensemble des points $(X_1, ..., X_n)$ *de* $\mathbb{R}^n$ *pour lequel l'hypothèse nulle* H_0 *est écartée au profit de l'hypothèse alternative* H_1*. On appelle aussi région critique du test et on la note généralement par* W*. Elle est définie par la relation*

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

Le complémentaire de la région critique est appelée région d'acceptation du test. Elle est notée par $\overline{W}$ *et est définie par*

$$P(\overline{W}|H_0) = 1 - \alpha.$$

1. L'indicatrice de W est appelée fonction critique du test. On note par

$$\delta(x_1, ..., x_n) := 1_w = \begin{cases} 1 \text{ si } (x_1, ..., x_n) \in W \\ 0 \text{ si } (x_1, ..., x_n) \notin W \end{cases}.$$

2. La construction d'un test est en fait la détermination de la région critique W.

   D'où en vertu de la relation remarque précédente, la nécessité de connaître la loi de probabilité de la variable de décision sous l'hypothèse H_0.

3. Puisque la puissance est $\pi = P[W|H_1]$ alors, pour son calcul, il est nécessaire de connaître la loi de probabilité de la variable de décision sous l'hypothèse H_1.

4. Certains auteurs s'intéressent au plus petit niveau de signification. Il s'appelle dimension du test. C'est le risque de première espèce maximum

$$\alpha := \sup_{\theta \in \Theta_0} \alpha(\theta) = \sup_{\theta \in \Theta_0} \pi(\theta).$$

Exemple 1.3.1

Soit $X \sim U(0, \theta)$, $\theta > 0$. On désire tester les deux hypothèses

$$\begin{cases} H_0 : \quad 3 \leq \theta \leq 4 \\ H_1 : \ \theta < 3 \text{ ou } \theta > 4 \end{cases}.$$

On a

$\Theta =]0, +\infty[$, $\Theta_0 = [3, 4]$ et $\Theta_1 =]0, 3[\cup]4, +\infty[$.

On suppose que la région d'acceptation du test est

$$\overline{W} = \{(x_1, ..., x_n) \in \mathbb{R}^n : \ 2.9 \leq Y_n \leq 4\}.$$

On rappelle que $Y_n := \max\{x_1, ..., x_n\}$ est l'estimateur de maximum de vraisemblance de θ et que sa fonction de répartition est définie sur $\mathbb{R}$ par

$$F_{Y_n}(x) := \begin{cases} 0 & \text{si } x < 0 \\ (\frac{x}{\theta})^n & \text{si } 0 < x < \theta \\ 1 & \text{si } x > \theta \end{cases},$$

la fonction puissance du test est définie sur $]0, +\infty[$ par

$$\begin{aligned} \pi(\theta) &= P(W) = P(Y_n < 2.94) + P(Y_n > 4) \\ &= F_{Y_n}(2.94) + 1 - F_{Y_n}(4), \end{aligned}$$

- si $\theta > 4$ alors $\pi(\theta) = (\frac{2.9}{\theta})^n + 1 - (\frac{4}{\theta})^n$,
- si $3 \leq \theta \leq 4$ alors $\pi(\theta) = (\frac{2.9}{\theta})^n$,
- si $\theta < 3$ alors $\pi(\theta) = \begin{cases} 1 & si\ \theta < 2.9 \\ (\frac{2.9}{\theta})^n & si\ 2.9 \leq \theta < 3 \end{cases}$,

la dimension du test est

$$\alpha = \sup_{3 \leq \theta \leq 4} \pi(\theta) = \pi(3) = (\frac{2.9}{3})^n,$$

pour un échantillon de taille 68 on trouve $\alpha \simeq 0.10$.

Propriétés d'un test statistique

Test sans biais On dit qu'un test est sans biais si sa function puissance reste supérieure ou égale à son niveau α : $1 - \beta \geq \alpha$.

Test convergent (ou consistant) Un test est dit convergent si sa puissance tend vers 1.

Test Uniformément plus puissant ($U.P.P.$)**[1]** Pour deux statistiques de test Φ et $\widetilde{\Phi}$ de niveau α, on dit que Φ est plus puissant que $\widetilde{\Phi}$ si la puissance de Φ est supérieure à la puissance de $\widetilde{\Phi}$.

Un test est dit uniformément plus puissant de niveau α s'il est uniformément plus puissant que tout autre test de niveau α.

Il n'existe pas toujours de test U.P.P.(α) mais nous verrons plus loin pour quels types d'hypothèses on peut en construire.

1.4 Protocole Statistique

Afin de réaliser un test statistique, on doit suivre généralement le schéma suivant

1. Formuler les hypothèses H_0 et H_1.
2. Fixer le niveau de signification α.
3. Détermination de la statique de test et de sa loi sous H_0.
4. Calcul de la valeur observée de la statistique de test.
5. Décision : rejet ou acceptation de H_0.[1]

1.4.1 P-valeur

En statistique la P-valeur est la probabilité pour un modèle statistique donne sous l'hypothèse nulle d'obtenir la même valeur ou une valeur encore plus extrême que celle observée sur l'échantillon. On va comparer un seuil de signification α et P-valeur pour accepter ou rejeter H_0 comme suit

- Si $P - valeur \leq \alpha$ on va rejeter l'hypothèse H_0.
- Si $P - valeur > \alpha$ on va accepter l'hypothèse H_0.

On peut alors interpréter la P-valeur comme le plus petit seuil de significativité pour lequel H_0 est acceptée.

1.5 Catégories des tests statistique

1.5.1 Tests paramétriques

Définition 1.5.1 *Un test paramétrique est un test de contrôler certaine hypothèse relative à un ou plusieurs paramètres comme (la moyenne, la variance ou la fréquence observé) d'une variable aléatoire de loi spécifiée ou non. Dans la plupart de ces tests basés sur la loi normale.*

parmi ces tests on cite les tests de conformité et les tests d'homogénéité.

1.5.2 Tests non paramétriques

Définition 1.5.2 *Un test non paramétrique est un test ne nécessitant pas d'hypothèse sur la forme des données, les statistiques de test sont alors remplacées par des statistiques ne dépendant pas des moyennes et variances des données initiales. Les tests non paramétriques les plus connus sont*

- *Le test d'indépendance ou d'association consiste à éprouver l'existence d'une liaison entre 2 variables. Les techniques utilisées diffèrent selon que les variables sont qualitatives nominales, ordinales ou quantitatives [14].*
- *Le test d'ajustement ou d'adéquation consiste à vérifier la compatibilité des données avec une distribution choisie a priori. Le test le plus utilisé dans cette optique est le test d'ajustement à la loi normale, qui permet ensuite d'appliquer un test paramétrique [14]. Dans ce mémoire on s'intéresse à cette famille du tests.*

1.6 Méthode de Neyman et Pearson

1.6.1 Tests entre hypothèses simples

Soit X une variable aléatoire de densité de probabilité f_θ où θ est un paramètre inconnu.

Il s'agit de tester [2]

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}.$$

Soit

$$L_\theta(x_1, ..., x_n) := \prod_{i=1}^{n} f_\theta(x_i),$$

la fonction de vraisemblance (ou bien la densité) de l'échantillon $(X_1, ..., X_n)$ de X.

On pose

$$L_i(x_1, ..., x_n) := L_{\theta_i}(x_1, ..., x_n),\ i = 0, 1,$$

le rapport

$$\frac{L_1(x_1, ..., x_n)}{L_0(x_1, ..., x_n)}$$

est appelé rapport de vraisemblance.

Lemme 1.6.1 *Neyman-Pearson (cas continu) [2]*

On suppose ici que la v.a X est continue. Un test δ_k est le test qui rejette H_0 au niveau de signification α. si et seulement si le rapport de vraisemblance est au moins k, où $k = k(\alpha) \geq 0$.

Si δ est un autre test tel que $\alpha(\delta) \leq \alpha(\delta_k)$, alors

$$\pi(\theta; \delta_k) \geq \pi(\theta; \delta),$$

c'est à dire δ_k est le plus puissant. En d'autres termes, la région critique optimale est

$$W_k := \left\{ (x_1, ..., x_n) \in \mathbb{R}^n : \frac{L_1(x_1, ..., x_n)}{L_0(x_1, ..., x_n)} \geq k \right\},$$

où k est telle que $P_0(W_k) = \alpha$ où $P_i(A) := P(A|H_i)$, $i = 0, 1$.

Le test δ_k est alors défini comme suit

$$\delta_k(x_1, ..., x_n) = \begin{cases} 1 \ si \ \frac{L_1(x_1,...,x_n)}{L_0(x_1,...,x_n)} \geq k \\ 0 \ si \ \frac{L_1(x_1,...,x_n)}{L_0(x_1,...,x_n)} < k \end{cases},$$

où k la solution de l'équation

$$P_0 \left\{ \frac{L_1(x_1, ..., x_n)}{L_0(x_1, ..., x_n)} \geq k \right\} = \alpha,$$

Autrement dit

$$\delta_k := E_0[\delta_k(X_1, ..., X_n)] = P_0(W) = \alpha,$$

où

$$E_i[X] := E[X|H_i], i = 0, 1.$$

Exemple 1.6.1 *Soit X une population normale d'espérance inconnue μ et de variance $\sigma^2 = 1$.*

On veut tester l'hypothèse $H_0 : \mu = 0$ contre l'hypothèse $H_1 : \mu = 1$.

Pour cela en prélève un échantillon de taille 9.

Quel est le test le plus puissant au niveau de signification 0.05 *?*

On a le rapport de vraisemblance

$$\begin{aligned}\frac{L_1(x_1,...,x_n)}{L_0(x_1,...,x_n)} &= \frac{\prod_{i=1}^{9}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x_i-1)^2\}}{\prod_{i=1}^{9}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x_i)^2\}} \\ &= \exp\{9(\overline{X}-\frac{1}{2})\},\end{aligned}$$

où $\overline{X}=\sum_{i=1}^{9}\frac{x_i}{9}$ *est la moyenne empirique. La région de rejet (critique) est donc*

$$\begin{aligned}W &= \left\{(x_1,...,x_n)\in\mathbb{R}^9 : \exp\{9(\overline{X}-\frac{1}{2})\}\geq k\right\} \\ &= \left\{(x_1,...,x_n)\in\mathbb{R}^9 : \overline{X}\geq k'\right\},\end{aligned}$$

où $k'=\frac{1}{2}+\frac{\log k}{9}$ *et telle que*

$$0.05 = P(W|\mu=0) = P_0(W) = P_0(\overline{X}\geq k')$$

Ou sous H_0, $\overline{X}\sim N(0,\frac{1}{9})$, *ainsi* $Z:=3\overline{X}\sim N(0,1)$, *d'où* $P(Z\geq 3k')=0.05$, *en d'autres termes* $\Phi(3k')=0.95$ *ou encore* $k'=\frac{1}{3}\Phi^{-1}(\alpha)$ *où* $\Phi^{-1}(\alpha)$ *désigne la fonction de quantile d'ordre* α *de la loi normale standard.*

De la table statistique de la loi normale (Gauss) on tire $\Phi^{-1}(0.95)=1.64$, *par conséquent* $k'=\frac{1.64}{3}=0.54$.

La région critique optimale est donc

$$W=\left\{(x_1,...,x_n)\in\mathbb{R}^n : \overline{X}\geq 0.54\right\},$$

Le test optimal (le plus puissant) est par conséquent

$$\delta(x_1, ..., x_n) = \begin{cases} 1 \ si \ 0.54 \geq k \\ 0 \ si \ 0.54 < k \end{cases}.$$

Calculant la puissance $1 - \beta$ du test δ

$$\begin{aligned} 1 - \beta = P(W|\mu = 1) &= P(\overline{X} \geq 0.54|\mu = 1) \\ &= 1 - P_1(\overline{X} < 0.54) \\ &= 1 - P_1(3(\overline{X} - 1) < 3(0.54 - 1)) \\ &= 1 - P(Z^* < -1.38), \text{où } Z^* \sim N(0, 1) \\ &= P(Z^* < 1.38) = 0.91, \end{aligned}$$

ainsi le risque de deuxième espèce est $\beta = 1 - 0.91 = 0.09$.

Lemme 1.6.2 *Lemme de Neyman-Pearson (cas discret) [2]*

Il arrive que pour certaines valeurs de α, il n'existe pas de constante k vérifiant l'équation $\alpha(\delta_k) = k$, ce qui ce passe surtout dans le cas où X est discrète. Dans ce cas, on modifie δ_k en définissant le δ_k^* (qui sera le plus puissant)

$$\delta_k^*(x_1, ..., x_n) = \begin{cases} 1 \ si \ \frac{L_1(x_1,...,x_n)}{L_0(x_1,...,x_n)} > k \\ p \ si \ \frac{L_1(x_1,...,x_n)}{L_0(x_1,...,x_n)} = k \\ 0 \ si \ \frac{L_1(x_1,...,x_n)}{L_0(x_1,...,x_n)} < k \end{cases}$$

où k et $0 < p < 1$ sont deux constantes définies, sous H_0, par l'équation

$$\alpha(\delta_k^*) = P_0\left\{\frac{L_1(x_1, ..., x_n)}{L_0(x_1, ..., x_n)} > k\right\} + pP_0\left\{\frac{L_1(x_1, ..., x_n)}{L_0(x_1, ..., x_n)} = k\right\} = \alpha.$$

On note que δ_k est appelé test du rapport de vraisemblance et δ_k^* est appelé test du rapport de vraisemblance randomisé.

Exemple 1.6.2 *On prélève un échantillon de taille 8, issu d'une v.a X de loi de poisson de paramètre $\lambda > 0$. Pour tester l'hypothèse $H_0 : \lambda = 1$ contre $H_1 : \lambda = 2$ au niveau de signification $\alpha = 0.1$.*

Déterminer le test le plus puissant et quelle est sa puissance ?

Rappelons que la loi de Poisson, de paramètre $\lambda > 0$, est définie par sa fonction de masse

$$P_\lambda(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}, x = 0, 1, 2, ...,$$

et sa fonction de répartition

$$P(X \leq x) = \sum_{k=0}^{x} P_\lambda(X = k) = e^{-\lambda}\sum_{k=0}^{x}\frac{\lambda^k}{k!}, x = 0, 1, 2, ...$$

Le rapport de vraisemblance qui correspond à ce test est

$$\frac{L_1(x_1, ..., x_8)}{L_0(x_1, ..., x_8)} = \frac{\prod_{i=1}^{8} P_{\lambda=2}(X = x_i)}{\prod_{i=1}^{8} P_{\lambda=1}(X = x_i)} = \frac{\prod_{i=1}^{8} \frac{2^{x_i}}{x_i!}e^{-2}}{\prod_{i=1}^{8} \frac{1^{x_i}}{x_i!}e^{-1}} = e^{-8}2^s,$$

où $S := x_1 + ... + x_8$. Alors le test statistique le plus puissant est

$$\delta(x_1, ..., x_8) = \begin{cases} 1 \ si \ e^{-8}2^S > k \\ p \ si \ e^{-8}2^S = k \\ 0 \ si \ e^{-8}2^S < k \end{cases},$$

où k et $0 < p < 1$ sont deux constantes définies sous H_0 par l'équation

$$P_0(e^{-8}2^S > k) + pP_0(e^{-8}2^S = k) = \alpha = 0.1,$$

En d'autres termes

$$\delta(x_1, ..., x_8) = \begin{cases} 1 \text{ si } S > c \\ p \text{ si } S = c \\ 0 \text{ si } S < c \end{cases},$$

où c et $0 < p < 1$ sont deux constantes définies sous H_0 par l'équation

$$P_0(S > c) + pP_0(S = c) = \alpha = 0.1,$$

cette équation peut être réécrite comme suit

$$P_0(S \leq c) = 0.90 + pP_0(S = c) > 0.90. \tag{1.1}$$

Sous H_0, l'échantillon de taille 8 provient d'une loi de Poisson de paramètre $\lambda = 1$, donc S est suit aussi une loi de Poisson de paramètre $8\lambda = 8$.

De la table statistique de la loi de Poisson, on remarque que la petite valeur de c vérifiant $P_0(S \leq c) > 0.90$ est $c = 12$, qui correspond à $P_0(S \leq 12) = 0.93$.

On en déduit que

$$\begin{aligned} P_0(S = 12) &= P_0(S \leq 12) - P_0(S \leq 11) \\ &= 0.93 - 0.88 = 0.05, \end{aligned}$$

ce qui implique, de l'équation (1.1), que

$$p = \frac{0.93 - 0.90}{0.05} = 0.60 \Longrightarrow 60\%,$$

En d'autres termes si $S = 12$ *on rejette* H_0 *avec une probabilité de* 60%*, et si* $S > 12$

en rejette H_0 *à 100%. Ainsi le test le plus puissant est*

$$\delta = \delta(x_1, ..., x_8) = \begin{cases} 1 \ si \ S > 12 \\ 0.60 \ si \ S = 12 \\ 0 \ si \ S < 12 \end{cases} .$$

Chapitre 2

Tests d'ajustement et applications

Les tests d'ajustement servent à tester si un échantillon est distribué selon une loi de probabilité donnée. Ils permettent de décider, avec un seuil d'erreur α spécifié, si les écarts présentés par l'échantillon par rapport aux valeurs théoriques attendues sont dûs au hasard ou sont au contraire significatifs. Alors, les hypothèse à tester dans ce cas sont de la forme suivantes

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

où la fonction de répartition F_0 est donnée.

2.1 Test du χ^2

2.1.1 Test d'ajustement du χ^2 pour une variable discrète

Soient O_i les effectifs observés pour chaque classe d'évènements et C_i les effectifs calculés qui sont les effectifs théoriques qu'on obtiendrait en appliquant la loi théorique.

Sous l'hypothèse H_0

H_0 : il y a adéquation de la distribution observée avec la distribution théorique,

on montre que la variable aléatoire E^2 définie par

$$E^2 = \sum_{i=1}^{n} \frac{(O_i - C_i)^2}{C_i},$$

suit une loi du χ^2 à $v = n - 1$ degrés de liberté. Voir [4]

La statistique E^2 s'appelle la distance du χ^2 entre le vecteur des valeurs observées et celui des valeurs calculées.

En utilisant les tables du χ^2, on détermine le quantile qui délimite la région d'acceptation au seuil α fixé. Si la distance E^2 est supérieure au quantile, on rejette l'hypothèse H_0.

Remarque 2.1.1

- Pour que ce test soit valide, il faut que $C_i \geqslant 5$ pour tout i. En particulier, on ne doit pas l'appliquer si une classe a un effectif nul.
- Dans le cas d'une variable continue, les données doivent être regroupées en classes et chaque classe est représentée par une valeur (par exemple, le milieu

de la classe). Le nombre n représente alors le nombre de classes.[3]

Exemple 2.1.1 *Loi uniforme*

Une statistique relative aux résultats du concours d'entrée à une grande école fait ressortir les répartitions des candidats et des admis selon la profession des parents.

Profession de candidats	Nombre de candidats	Nombre d'admis
Fonctionnaires et assimilés	2245	181
Commerce, industrie	986	84
Professions libérales	574	47
Propriétaires rentiers	422	38
Propriétaires agricoles	288	12
Artisans, petits commerçants	211	19
Banque, assurance	208	17
Total	4934	398

TAB. 2.1 – Résultats du concours d'entrée à l'école supérieure.

Problème : Tester l'hypothèse (risque $\alpha = 0,05$) selon laquelle la profession des parents n'a pas d'influence sur l'accès à cette grande école.

Il s'agit du test d'ajustement d'une distribution théorique, on pose les hypothèses

- H_0 : "la profession des parents n'a pas d'influence sur l'accès à cette grande école"

c'est à dire la proportion des admis est constante pour toutes les professions soit $p = \frac{398}{4934} \simeq 0.0806$

- H_1 : " la profession des parents influe sur l'accès à cette grande école "

Sous H_0, le nombre d'admis pour la i-ième profession est C_i.

i	n_i	O_i effectif observé	C_i effectif théorique	$\frac{(O_i-C_i)^2}{C_i}$
1	2245	181	$\frac{2245\times398}{4934} \simeq 181.09$	$\frac{(181-181.09)^2}{181.09} \simeq 0$
2	986	84	$\frac{986\times398}{4934} \simeq 79.53$	$\frac{(84-79.53)^2}{79.53} \simeq 0.25$
3	574	47	$\frac{574\times398}{4934} \simeq 46.30$	$\frac{(47-46.30)^2}{46.30} \simeq 0$
4	422	38	$\frac{422\times398}{4934} \simeq 34.04$	$\frac{(38-34.04)^2}{34.04} \simeq 0.46$
5	288	12	$\frac{288\times398}{4934} \simeq 23.23$	$\frac{(12-23.23)^2}{23.23} \simeq 5.42$
6	211	19	$\frac{211\times398}{4934} \simeq 17.02$	$\frac{(19-17.02)^2}{17.02} \simeq 0.23$
7	208	17	$\frac{208\times398}{4934} \simeq 16.77$	$\frac{(17-17.02)^2}{17.02} \simeq 0$
Total	4934	398	397.98	6.36

TAB. 2.2 – Table des valeurs empiriques.

Le χ^2 observé vaut 6.36. Le nombre de degrés de liberté est 7-1 = 6. La table de χ^2 fournit $\chi^2_{6;0.95} = 12.59$ donc χ^2 observé $< \chi^2_{6;0.95}$. On ne rejette pas H_0, ce que la profession des parents n'a pas d'influence sur l'accès à cette grande école.

2.1.2 Test d'ajustement du χ^2 pour une variable continue

Si l'on pose la question de savoir si une variable aléatoire X suit ou non la loi normale $N(0,1)$, on peut se ramener au problème précédent en discrétisant la variable, c'est-à-dire que l'on fait une partitionnie de l'ensemble $\mathbb{R}$ de toutes les valeurs possibles de X formée de r intervalles successifs sans point commun

$$]-\infty, a_1[,]a_1, a_2], ...,]a_{r-1}, +\infty[.$$

Si l'on a observé un n-échantillon de valeurs de X soient $x_1, x_2, ..., x_n$, on résume ces observations en $(N_1, ..., N_r)$ où N_i désigne le nombre des x_i qui sont inférieurs à a_1, N_2 le nombre de ceux qui tombent entre a_1 (non compris) et $a_2, ...$

Sous l'hypothèse

$$H_0 : X \sim N(0, 1),$$

les probabilités p_j pour que X tombe dans chacun des r intervalles $I_j =]a_{j-1}, a_j]$ peuvent être calculées

$$p_j = \int_{a_{j-1}}^{a_j} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx,$$

et on voit donc comment se ramener au problème du paragraphe précédent pour toute loi continue dont la densité est complètement spéciée.[4]

2.2 Test de Kolmogorov-Smirnov

C'est le plus populaire parmi les tests d'ajustement qui sont basés sur la fonction de répartition empirique (F_n). Il a été proposé par Andreï N. Kolmogorov en 1933 et étendu par Vladimir I. Smirnov en 1939.[5]

2.2.1 Statistique du test

La distance utilisée pour définir la statistique D_n de ce test est celle de la norme uniforme. La statistique de Kolmogorov-Smirnov est alors définie par

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|,$$

où F_n est la fonction de répartition empirique.

Remarque 2.2.1 *Puisque* $0 \leq F_0(x) \leq 1$ *et* $0 \leq F_n(x) \leq 1$, *alors* $0 \leq D_n \leq 1$.

Pour calculer les valeurs de la statistique D_n, *il suffit d'évaluer la différence entre* F_n *et* F_0 *aux points* $x_{(i)}$ *comme l'indique la proposition suivante*

Proposition 2.2.1 *[6]*

La statistique de Kolmogorov-Smirnov s'écrit comme suit

$$D_n := \max\{\max_{1\leq i\leq n}[\frac{i}{n} - F_0(x_{(i)})], \max_{1\leq i\leq n}[F_0(x_{(i)}) - \frac{i-1}{n}], 0\}.$$

Preuve. [6]

La statistique D_n peut s'écrire $D_n = \max(D_n^+, D_n^-)$, avec

$$D_n^+ := \sup_{x\in\mathbb{R}}(F_n(x) - F_0(x)) \text{ et } D_n^- := \sup_{x\in\mathbb{R}}(F_0(x) - F_n(x)),$$

on a

$$D_n^+ = \sup_{x\in\mathbb{R}}(F_n(x) - F_0(x)) = \max_{0\leq i\leq n} \sup_{x_{(i)}\leq x\leq x_{(i+1)}} (F_n(x) - F_0(x)),$$

on définit $x_{(0)} = -\infty$ et $x_{(n+1)} = +\infty$, on peut écrire $F_n(x) = i/n$ pour $x_{(i)} \leq x < x_{(i+1)}, i = 0, 1, ..., n$, alors on a

$$D_n^+ = \max_{0\leq i\leq n} \sup_{x(i)\leq x\leq x_{(i+1)}} (\frac{i}{n} - F_0(x)) = \max_{0\leq i\leq n}(\frac{i}{n} - \inf_{x(i)\leq x\leq x_{(i+1)}} F_0(x)),$$

où F_0 est une fonction croissante, d'où

$$D_n^+ = \max_{0\leq i\leq n}(\frac{i}{n} - F_0(x_{(i)})) = \max\{\max_{1\leq i\leq n}[\frac{i}{n} - F_0(x_{(i)})], 0\}. \tag{2.1}$$

Par le même principe, on montre le résultat relatif à D_n^-. ■

Remarque 2.2.2 *[5]*

Dans le cas où F_0 est continue, les lois de D_n, D_n^+ et D_n^- sont indépendantes de F_0. En effet, si F_0 est continue, les v.a's $F_0(X_{(i)}), i = 1, 2, ..., n$, sont uniformes sur [0,1], c'est-à-dire $F_0(X_{(i)}) \stackrel{l}{=} U_{(i)}$, pour $i = 1, 2, ..., n$ indépendamment de F_0. Par conséquent, D_n, D_n^+ et D_n^- ont des distributions indépendantes de F_0.

Par le changement de variable $y = F_0(x)$, on peut écrire

$$D_n = \sup_{0 \leq y \leq 1} |G_n(y) - y|,$$

où G_n est la fonction de répartition empirique uniforme.

Thèorème 2.2.1 *[6]*

Pour tout $t > 0$, on a

$$\lim_{n \to \infty} P(\sqrt{n} D_n < t) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2 t^2).$$

Thèorème 2.2.2 *[6]*

Dans le cas où F_0 est continue, on a, pour tout réel t et $n \geq 1$,

$$P(D_n < t) = \begin{cases} 0 & si\ t \leq \frac{1}{2n} \\ \int_{1/n-t}^{t} \int_{5/6n-t}^{t-1/6n} ... \int_{1-t}^{(n-1)/n+t} f(u_1, u_2, ..., u_n) du_1 du_2 ... du_n & si\ \frac{1}{2n} < t < 1 \\ 1 & si\ t \geq 1 \end{cases},$$

où

$$f(u_1, u_2, ..., u_n) = n! 1_{(0 < u_1 < u_2 < ... < u_n < 1)}.$$

Preuve. [6]

Pour loi uniforme on a

$$D_n^+ = \max\{\max_{1\leq i\leq n}(\frac{i}{n} - x_{(i)}), 0\},$$

et pour $0 < c < 1$ on a

$$\begin{aligned} P(D_n^+ < c) &= P[\max(\frac{i}{n} - X_{(i)}) < c] \\ &= P(\frac{i}{n} - X_{(i)} < c \text{ pour toute } i = 1, 2, ..., n) \\ &= P(X_{(i)} > \frac{i}{n} - c \text{ pour toute } i = 1, 2, ..., n) \\ &= \int_{1-c}^{\infty}\int_{(n-1)/n-c}^{\infty} ... \int_{2/n-c}^{\infty}\int_{1/n-c}^{\infty} f(x_1, x_2, ..., x_n)dx_1...dx_n, \end{aligned}$$

où

$$f(x_1, x_2, ..., x_n) = \begin{cases} n! & 0 < x_1 < x_2 < ... < x_n < 1 \\ 0 & sinon \end{cases}.$$

■

2.2.2 Région critique du test

La région critique du test pour la statistique de Kolmogorov-Smirnov est définie par

$$D_n > KS_{n,1-\alpha},$$

où la valeur $(KS_{n,1-\alpha})$ étant donnée par la table de Komogorov-Smirnov.

- Une valeur élevée de D_n est une indication que la distribution de l'échantillon s'éloigne sensiblement de la distribution de référence $F_0(x)$, et qu'il est donc

peu probable que H_0 soit correcte. Plus précisément

$$P(\sup|F_n(x) - F_0(x)| > \frac{c}{\sqrt{n}}) \xrightarrow[n\longrightarrow+\infty]{} \alpha(c) = 2\sum_{r=1}^{+\infty}(-1)^{r-1}\exp(-2r^2c^2),$$

pour toute constante $c > 0$. Le terme (c) vaut 0.05 pour $c = 1.36$. Pour $n > 100$, la valeur critique du test est approximativement de la forme $\frac{c}{\sqrt{n}}$. Les valeurs de c en fonction de valeurs usuelles de α sont données dans le tableau suivant

α	0.20	0.10	0.05	0.02	0.01
c	1.073	1.223	1.358	1.518	1.629

TAB. 2.3 – Valeurs de c pour calculer la valeur critique du test.

On rejette H_0 si

$$D_n > \frac{c}{\sqrt{n}}.$$

Exemple 2.2.1

Pour population Ω, on veut étudier la conformité de la distribution d'une v.a continue (X) à une distribution normale, on dispose pour cela un échantillon de taille $n = 30$ observations suivants

$$X = (14, 14, 18, 17, 16, 17, 2, 6, 9, 13, 12, 0, 6, 2, 20,$$
$$21, 28, 30, 21, 32, 10, 20, 23, 22, 10, 13, 11, 13, 13, 12),$$

on va tester les hypothèses suivantes

$$\begin{cases} H_0 : \text{la variable } X \text{ suit une loi normale} \\ H_1 : \text{la variable } X \text{ ne suit pas la loi normale} \end{cases},$$

calculs du test

1. trier des données brutes en ordre croissant $(x_{(i)})_{i=1}^n, n = 30$;
2. centrage et réduction des valeurs de X;

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S}, \text{où } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{445}{30} = 14.83 \text{ et } S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} = 7.8743,$$

3. trouver les valeurs de Z correspondantes avec le tableau de loi normale ;
4. trouver les valeurs maximale de

$$\max_i[\frac{i}{n} - F_0(x_{(i)})] = 0.10881, \ \max[F_0(x_{(i)}) - \frac{i-1}{n}] = 0.07547$$

   donc

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| = 0.10881;$$

5. on a pour $\alpha = 0.05$ et $n = 30$

$$D_n = 0.10881 < KS_{n,1-\alpha} = 0.2417.$$

Alors au risque $\alpha = 0.05$, on accepte l'hypothèse H_0.

Code R :

X=c(14,14,18,17,16,17,2,6,9,13,12,0,6,2,20,21,28,30,21,32,10,20,23,22,10,13,11,13,13,12)

ks.test(X, pnorm, mean=mean(X), sd=sd(X))

One-sample Kolmogorov-Smirnov test

data : X

$D = 0.10881$, p-value $= 0.8696$

alternative hypothesis : two-sided

Commentaire : On remarque que $p - value > \alpha$ $(0.8696 > 0.05)$, alors au risque $\alpha = 0.05$ on accepte l'hypothèse H_0 (la variable X suit une loi normale).

2.3 Test de Lilliefors

Ce test est une variante du test de Kolmogorov-Smirnov, sous l'hypothèse de normalité (à cherche à tester $H_0 : X \sim N(\mu, \sigma^2)$), où les paramètres de la loi sont estimés à partir des données.

alors au niveau de signification $\alpha = 0.05$, on peut ajuster les observation à une distribution normale. [8]

2.3.1 Statistique du test

$$L = \max_{1 \leq i \leq n} (F_i - \frac{i-1}{n}, \frac{i}{n} - F_i),$$

où F_i est la fréquence théorique de la loi de répartition normale centrée et réduite associée à la valeur standardisée.

$$Z_{(i)} = \frac{x - \overline{X}}{S}.$$

2.3.2 Région critique

La région critique du test pour la statistique L est définie par

$$R.C : L > L_{crit}.$$

La table des valeurs critiques L_{crit} pour les petites valeurs de n et différentes valeurs de α doivent être utilisées. Lorsque les effectifs sont élevés, typiquement $n \geq 30$, il est possible d'approcher la valeur critique à l'aide de formules simples

α	L_{crit}
0.10	$\frac{0.885}{\sqrt{n}}$
0.05	$\frac{0.886}{\sqrt{n}}$
0.01	$\frac{1.031}{\sqrt{n}}$

TAB. 2.4 – Valeurs critiques du test Lilliefors.

Exemple 2.3.1 *On prend le même exemple 2.2.1.*

Calculs du test

1. trier des données brutes en ordre croissant $(x_{(i)})_{i=1}^{n}, n = 30$;
2. centrage et réduction des valeur de X;

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S}, \text{où } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{445}{30} = 14.83 \text{ et } S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} = 7.8743,$$

3. trouver les valeurs de Z correspondantes avec le tableau de loi normale ;

4. utilisons la fonction de répartition de la loi normale centrée et réduite pour extraire les fréquences théorique F_i;

5. que nous opposons aux fréquences empiriques pour obtenir la statistique D du test, en calculant

$$L = \max_{1 \leq i \leq n} (F_i - \frac{i-1}{n}, \frac{i}{n} - F_i) = 0.10881\,;$$

6. comparer au seuil critique $L_{crit} = 0.16176$ lue dans la table de Lilliefors à une risque $\alpha = 0.05$, pour $n = 30$

$$L < L_{crit}.$$

Alors au risque $\alpha = 0.05$, on accepte l'hypothèse H_0

Code R :

X=c(14,14,18,17,16,17,2,6,9,13,12,0,6,2,20,21,28,30,21,32,10,20,23,22,10,13,11,13,13,12)

lillie.test(X)

Lilliefors (Kolmogorov-Smirnov) normality test

data : X

$D = 0.10881$, p-value $= 0.4853$

Commentaire : On remarque que $p - value > \alpha$ $(0.4853 > 0.05)$

donc, pour $\alpha = 0.05$, on peut accepter l'hypothèse H_0 (la variable X suit une loi normale).

2.4 Test de Cramer-von Mises

Le test était développé par Harald Cramer et Richard E. von Mises (1928-1930). Voir [11]

2.4.1 Statistique du test

La statistique de Cramer-von Mises est définie par

$$W_n^2 := n \int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 dF_0(x),$$

ou par

$$W_n^2 := \sum_{i=1}^{n} (F_0(x_{(i)}) - \frac{2i-1}{2n})^2 + \frac{1}{12n}.$$

2.4.2 Région critique

On rejette H_0 si

$$W_n^2 \geq W_{crit}^2.$$

Pour un niveau α donné et pour $n = 30$, les valeurs critiques sont résumées dans le tableau suivant

α	W_{crit}^2
0.10	0.172
0.05	0.218
0.01	0.33

TAB. 2.5 – Valeurs critiques du test Cramer-von Mises.

Exemple 2.4.1 *On prend le même exemple 2.2.1.*

Calculs du test

1. trier des données brutes en ordre croissant $(x_{(i)})_{i=1}^{n}, n = 30$;
2. centrage et réduction des valeur de X;

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S}, \text{où } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{445}{30} = 14.83 \text{ et } S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} = 7.8743\,;$$

3. trouver les valeurs de Z correspondantes avec le tableau de loi normale ;
4. utilisons la fonction de répartition de la loi normale centrée et réduite pour extraire les fréquences théorique $F_0(x_{(i)})$;
5. calculer la statistique

$$W_n^2 = \sum_{i=1}^{n}(F_0(x_{(i)}) - \frac{2i-1}{2n})^2 + \frac{1}{12n} = 0.041701 < W_{crit}^2 = 0.218.$$

Alors au risque $\alpha = 0.05$, on accepte l'hypothèse H_0

Code R :

X=c(14,14,18,17,16,17,2,6,9,13,12,0,6,2,20,21,28,30,21,32,10,20,23,22,10,13,11,13,13,12)

cvm.test(X)

Cramer-von Mises normality test

data : X

$W = 0.041701$, p-value = 0.6374

alternative hypothesis : two-sided

Commentaire : On remarque que $p-value > \alpha$ $(0.6374 > 0.05)$, donc on accepte l'hypothèse de la normalité (H_0) au seuil de risque $\alpha = 5\%$.

2.5 Test d'Anderson-Darling

2.5.1 Statistique du test

Ce test est sensible aux données en queues de distribution. Sa statistique A est

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[(\log(F_i) + \log(1 - F_{n-i+1}))^2].$$

Une correction a été proposé par Stephens

$$A^* = A(1 + \frac{0.75}{n} + \frac{2.25}{n^2}),$$

F_i étant la même fonction que le test de Kolmogorov-Smirnov. [10]

2.5.2 Région critique

L'hypothèse de normalité est rejetée si

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

Les valeurs critiques A_{crit} pour différents niveaux de risques sont résumées dans le tableau suivant

α	A_{crit}
0.10	0.631
0.05	0.752
0.01	1.035

TAB. 2.6 – Valeurs critiques du test d'Aderson-Darling.

Exemple 2.5.1 *On prend le même exemple 2.2.1.*

Calculs du test

1. trier des données brutes en ordre croissant $(x_{(i)})_{i=1}^{n}, n = 30$;
2. centrage et réduction des valeurs de X;

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S}, \text{où } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{445}{30} = 14.83 \text{ et } S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} = 7.8743\,;$$

3. trouver les valeurs de Z correspondantes avec le tableau de loi normale ;
4. utilisons la fonction de répartition de la loi normale centrée et réduite pour extraire les fréquences théoriques F_i,et calculer $\log(F_i)$;
5. former F_{n-i+1} puis en déduire $\log(1 - F_{n-i+1})$;
6. calculer la statistique

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)(\log(F_i) + \log(1 - F_{n-i+1}))^2] = 0.26532\,;$$

7. comparer au seuil critique $A_{crit} = 0.752$ une risque $\alpha = 0.05$

$$A < A_{crit}.$$

Alors au risque $\alpha = 0.05$, on accepte l'hypothèse H_0

Code R :

X=c(14,14,18,17,16,17,2,6,9,13,12,0,6,2,20,21,28,30,21,32,10,20,23,22,10,13,11,13,13,12)

ad.test(X)

Anderson-Darling normality test

data : X

$A = 0.26532$, p-value $= 0.669$

Commentaire : On remarque que $p - value > \alpha$

on constate alors, qu'au risque $\alpha = 5\%$ l'hypothèse H_0 acceptée c'est à dire la variable aléatoire$X$ suit une loi normale.

2.6 Test de Shapiro-Wilk

Le test de Shapiro-Wilk est basé sur la statistique W. En comparaison des autres tests, il est particulièrement puissant pour les petits effectifs ($n \leq 50$). La statistique du test s'écrit

$$W = \frac{[\sum_{i=1}^{n} a_i x_i]^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

où

- $x_{(i)}$ correspond à la série des données triées,
- les α_i sont des constantes générées à partir de la moyenne et de la matrice de covariance des quantiles d'un échantillon de taille n suivant la loi normale. ces constantes sont fournies dans des tables spécifiques.

La statistique W peut donc être interprétée comme le coefficient de détermination (le carré du coefficient de corrélation) entre la série des quantiles générées à partir la loi normale et les quantiles empiriques obtenues à partir des données, plus la compatibilité avec la loi normale est crédible.

Région critique

La région critique, rejet de la normalité, s'écrit

$$R.C : W < W_{crit}.$$

Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk.

Exemple 2.6.1 *On prend le même exemple 2.2.1.*

Calculs du test

1. Trier les données xi, nous obtenons la série $x_{(i)}$;
2. calculer les quantités

$$(x_{(n-i+1)} - x_{(i)}), i = 1, ..., [\frac{n}{2}];$$

3. lire dans la table pour $n = 30$ et $i = \overline{1.15}$ les valeur de coefficient α_i;
4. calculer la statistique

$$W = \frac{[\sum_{i=1}^{n} \alpha_i x_i]^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = 0.97656;$$

5. pour une risque $\alpha = 0.05$, le seuil critique lue dans la table de Shapiro-Wilk

$$W_{crit} = 0.927;$$

6. comparer entre W et W_{crit}

$$W > W_{crit}.$$

Alors au risque $\alpha = 0.05$, on accepte l'hypothèse H_0

Code R :

X=c(14,14,18,17,16,17,2,6,9,13,12,0,6,2,20,21,28,30,21,32,10,20,23,22,10,13,11,13,13,12)

shapiro.test(X)

Shapiro-Wilk normality test

$W = 0.97656$, $p - value = 0.7285$

Commentaire : On remarque que $p - value > \alpha$

donc on accepte l'hypothèse H_0 au risque $\alpha = 5\%$ (la v.a X suit une loi normale).

Conclusion

Le but de ce mémoire est d'appliquer les tests d'ajustement pour un échantillon statistique. Pour effectuer ce processus, premièrement on a défini les notions de base liés aux tests statistiques.

Deuxièmement, on a étudier les test statistique qui serts à vérifier qu'une variable aléatoire mesurée dans une population suit une loi de probabilité théorique donnée(les tests d'ajustement), en particulier on a étudier le test χ^2 avec ces deux types : le test d' ajustement du khi-deux dans le cas continu et discret et le test de Kolmogorov-Smirnov,... aussi deux autres types de tests d'justement qui ont pour but de tester la normalité ont été donné. puis à l'aide des exemples et sous le logiciel R, on a appliqué tous les tests mentionnés.

Le travail peut être s'enrichi par l'étude du test d'justement au cas multivarié.

Bibliographie

[1] G. Stoltz et V. Rivoirard, (2012). Statistique en Action. Institut de mathématiques de Toulouse.

[2] Abdelhakim NECIR, (2021). Cours de première master Université Mohamed Khider, Biskra.

[3] B. Desgraupes, (2018). Université Paris ouest nanterre la défense U.F.R. segmi.

[4] Ecole de Commerce International Dunkerque, (2015/2016) tests paramétriques et non paramétriques.

[5] GUESMIA Nour El Houda, (2018). Tests d'ajustement à une distribution basés sur la fonction de répartition empirique (mémoire). Université Mohamed Khider de biskra.

[6] Gibbons, J. D. & Chakraborti, S. (2010). Nonparametric statistical inference. CRC Press.

[7] FRANÇOIS Éthier, (2011). À propos de divers tests statistiques pour l'égalité de lois. Université du Québec à trois-rivières.

[8] Achour Chams, (2021). Tests de normalité et applications (mémoire). Université Mohamed Khider de biskra.

[9] Ricco Rakotomalala, R. (2008). Tests de normalité (techniques empiriques et tests statistiques). Université Lumière Lyon.

[10] Thibaut Martini, (2010). Détermination d'une méthode de calcul de capabilités avec des lois non gaussiennes. Université du Strasbourg.

[11] Pablo Martínez-Camblor, (2014). Cramer-Von Mises Statistic for Repeated Measures. Revista Colombiana de Estadística.

[12] Mouchiroud, D. (2003). Mathématique : "outils pour la biologie." Deug SV1 UCBL

[13] Akakpo, N. (7 septembre 2017). Tests statistiques. Master 1 mathématiques et applications université pierre et marie curie.

[14] Jean-Jacques Ruch, (2012-2013). Statistique : tests d'hypothéses. Préparation à l'Agrégation Bordeaux.

Annexe A : Logiciel R

2.7 Qu'est-ce-que le langage R ?

- Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèses, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle. R a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets.
- Il a été initialement créé, en 1996, par Robert Gentleman et Ross Ihaka du département de statistique de l'Université d'Auckland en Nouvelle Zélande. Depuis 1997, il s'est formé une équipe "R Core Team" qui développe R. Il est conçu pour pouvoir être utilisé avec les systèmes d'exploitation Unix, Linux, Windows et MacOS. L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.
- Un élément clé dans la mission de développement de R est le Compréhensive R

Archive Network (CRAN) qui est un ensemble de sites qui fournit tout ce qui est nécessaire à la distribution de R, ses extensions, sa documentation, ses fichiers sources et ses fichiers binaires. Le site maître du CRAN est situé en Autriche à Vienne, on peut y accéder par l'URL : "http ://cran.r-project.org/". Les autres sites du CRAN, appelés sites miroirs, sont répandus partout dans le monde.

Annexe B : Abréviations et Notations

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

Ω	Ensemple de population.
Θ	Ensemble des valeurs de θ.
H_0	Hypothèse nulle.
H_1	Hypothèse alternative.
α	Risque de premier espèce.
X	Variable aléatoire.
$\overline{X}$	Moyenne empirique.
$N(\mu, \sigma^2)$	Loi normale de paramètres μ et σ^2.
$N(0, 1)$	Loi normale centrée réduite.
$v.a$	Variable aléatoire.
S	Écart type.
$R.C$	Région critique.
D_n	Statistique de Kolmogorov-Smirnov.
D_{crit}	Valeur critique de Kolmogorov-Smirnov.

W_n^2	Statistique de Cramer-von Mises.
W_{crit}^2	Valeur critique de Cramer-von Mises.
L	Statistique de Lilliefors.
L_{crit}	Valeur critique de Lilliefors.
A	Statistique d'Anderson-Darling.
A_{crit}	Valeur critique d'Anderson-Darling.
W	Statistique de Shapiro-Wilk.
W_{crit}	Valeur critique de Shapiro-Wilk.