République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

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

DÉPARTEMENT DE MATHÉMATIQUES

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

MASTER en Mathématiques

Option : **Statistique**

Par

Rahmouni Yasmine

Titre :

Tests de normalité

Membres du Comité d'Examen :

Dr.	Berkane Hassiba	UMKB	Président
Dr.	Chine Amel	UMKB	Encadreur
Dr.	Dhiabi Samra	UMKB	Examinateur

Juin 2019

Dédicace

Je dédie ce humble travail à mes parents Lihlali et Saida sont trés chers pour me soutenir.

A mes frères Fares, Ammar

A mes soeurs Faiza, Nerdjes

je vous aime énormément.

A toutes ma famille et mes amies : Merieme, Manal, Samia,

A mon oncle Taher et sa femme Amina

vous à tous ceux qui m'encouragé

et à tous mes collègues de ma promotion de mathématiques.

RMERCIEMENTS

J'exprime d'abord mes profonds remerciements à "$\mathcal{ALLAH}$" qui m'a donné le courage et la volonté d'achever ce travail malgré les difficultés que nous avons rencontrées au cours des derniers mois.

Je tiens a exprimer toute ma reconnaissance a mon encadreur de mémoire, Dr **Amel Chine** je le remercie de m'avoir encadré orienté, aidé et conseillé.

Je tiens à remercier spécialement les memebres du Jury Dr **Berkane Hassiba** et, Dr **Dhaibi Samra** ont bien acceptès de participer et d'examiner mon modèste travail et pour ses conseils et pour toute l'aide qu'elles m'a apporté durant mes années universitaires .

J'adresse mes sincères remerciements à tous les professeurs du département de MATH présidé par Monsieur **Mokhtar Hafayedh,**

En fin, je remercie mes très chers parents et toutes ma famille et mes amis qui ont toujours été là pour moi leurs soutien inconditionnel et leurs encouragements on été d'une grande aide à tous ces intervenants, je présente mes remerciements, mon respect et ma gratitude.

Table des matières

Table des figures

Liste des tableaux

Introduction

En statistique, un test est une méthode de travail de vérification lors de la programmation d'un logiciel en laissant entrevoir les subtilités intervenant dans la traduction d'un problème de test qui illustrant la démarche conduisant à formuler les tests statistiques, sous la forme des tests de signification, il est nécessaire de préciser qu'un test de signification traite d'une certaine hypothèse statistique, c'est à dire d'un énoncé spécifiant la loi de distribution d'une population statistique souvent il s'agit d'un énoncé relatif à la valeur d'un ou plusieurs paramètres. Il existe de nombreux tests pour vérifier qu'un échantillon suit ou non une loi de probabilité donné, ce type de test s'appelle tests d'adéquation, Il s'agit de modélisation, parmi les tests d'adéquation (ajustement) la conformité à la loi normale (loi Gaussienne, loi de Laplace Gauss).

La distribution le plus utilisée dans l'analyse statistique est la distribution normale, par fois appel la distribution Gaussienne. Dans ce mémoire nous avons présenté les techniques statistiques et les méthodes destinées à évaluer la compatibilité d'une distribution empirique avec la loi normale.

En premier chapitre, nous avons rappelé les généralités sur la loi normale et tests d'hypothèses qui est un procédé d'inférence permettant de contrôler à partir de l'étude d'un ou plusieurs échantillons aléatoires dans ce support le principe des tests d'hypothèses est de

poser une hypothèse de travail qui est définie par :

$$\begin{cases} H_0 : \text{Loi de } X \text{ est une normale.} \\ \text{contre} \\ H_1 : \text{Loi de } X \text{ n}'\text{est pas une normale.} \end{cases}$$

Et on suite on distinguera deux classes de tests sont les tests paramétriques et les tests non paramétriques.

En deuxième chapitre, nous avons présenté deux méthodes pour vérifier la normalité des données. D'abord, nous avons expliqué la méthode graphique, un examen préalable des données à l'aide de graphique peut déjà permettre de visualiser si la distribution empirique suit une loi normale. Il s'agit donc de s'assurer que les variables continues sont distribuées selon la loi normale, si cela est le cas, les tests d'hypothèses classiques sont applicables, les graphes les très populaire sont l'histogramme de fréquence, la boite à moustache, le graphe quantile quantile (Q-Q plot) et la droite de Henry. En suite il y a la méthode théorique, nous détaillerons plusieurs tests statistiques comme test de Kolmogorov-Smirnov, test de Shapiro-Wilk, test de Lilliefors, test de d'Anderson-Darling, test de Cramer-von Mises, test de Jarque-Bera et test de D'Agostino, nous avons prendre un exemple sur un échantillon pour vérifier la normalité de ce dernier puis implémentés dans logiciel de statistique R.

Chapitre 1

Loi normale et tests d'hypothèses

Dans ce premier chapitre, on va d'abord présenter dans la première section une loi importante en statistique et probabilité c'est la loi normale au vu de ses caractéristiques, telle que le théorème centrale limite puis on présente dans la deuxième section les tests d'hypothèses (définition, types, exemples,...).

1.1 Loi normale

1.1.1 Quelque définition et propriété

Définition 1.1.1 *Soit X une variable aléatoire continue. On dit que X suit une loi normale $\mathcal{N}(\mu, \sigma)$ (ou loi gaussienne) si sa fonction de densité $f(x)$, pour tout nombre réel x, est définie par :*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); \quad x \in \mathbb{R},$$

avec μ et σ^2 sont La moyenne et la variance de X respectivement.

Propriété 1.1.1 *Si* $X \sim \mathcal{N}(\mu; \sigma^2)$ *alors :*

1. *La fonction de répartition de* X *est :*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \ .$$

2. *La moyenne de* X *est égale à* $E(X) = \mu$,

3. *La variance de* X *est égale à* $Var(X) = \sigma^2$.

1.1.2 Loi normale standard

Lorsque $\mu = 0$ et $\sigma = 1$ la loi normale $\mathcal{N}(\mu; \sigma^2)$ est dite loi centrée réduite ou loi normale standard, et on la note $Z \sim \mathcal{N}(0; 1)$, sa fonction de densité est donnée par :

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad ; \ Z \in \mathbb{R} \ , \tag{1.1}$$

et sa fonction de répartition, notée $\Phi(z)$, définie par :

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \ , \quad Z \in \mathbb{R}.$$

Propriété 1.1.2 *Par symétrie* $P(Z \leq -z) = 1 - P(Z \leq z)$ *; c'est à dire pour tout* $Z \in \mathbb{R}$ *;*

$\Phi(-z) = 1 - \Phi(z)$; *les valeur de* $\Phi(z)$ *sont données dans la table de loi normale centrée réduite* $\mathcal{N}(0; 1)$ *[6]*

Exemple 1.1.1

1. $\Phi(0) = P(Z \leq 0) = 0.5$;

2. $\Phi(1,96) = P(Z \leq 1,96) = 0,975$;

3. $\Phi(-1,96) = P(Z \leq -1,96) = 1 - P(Z \leq 1,96) = 1 - \Phi(1,96) = 1 - 0,975 = 0,025$;

Le calcul des probabilité d'une v.a X suit une loi normale $\mathcal{N}(\mu; \sigma^2)$ se ramène toujours a

celui de la loi normale $\mathcal{N}(0;1)$ comme indique le théorème (1.1.1)

Théorème 1.1.1 *Soit X une v.a suit la loi normale d"espérance μ et de variance σ^2 alors :*

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0;1). \tag{1.2}$$

par conséquent

$$F(x) = P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}), \quad x \in \mathbb{R}\ .$$

Exemple 1.1.2 *Soit X une v.a de loi $\mathcal{N}(2;9)$, on va calculer :*

1. $P(4 < X \leq 5)$.
2. $P(X \leq 2,8)$.

   Utilisant la table de la loi normale centrée réduite [6], on obtient les calcules suivants :

1. $P(4 < X \leq 5) = \Phi(\frac{5-2}{3}) - \Phi\left(\frac{4-2}{3}\right) = \Phi(1) - \Phi(0,67) = 0,8413 - 0,7486 = 0,0927.$
2. $P(X \leq 2,8) = \Phi\left(\frac{2,8-2}{3}\right) = \Phi(0,27) = 0,6064.$

1.1.3 Théorème centrale limite

Théorème 1.1.2 *Soit $\{X_n\}_{n\in\mathbb{N}}$ une suite de v.a indépendantes de même loi admettant une moyenne μ et une variance σ^2. Alors la suite $\frac{\overline{X_n}-\mu}{\sigma/\sqrt{2\pi}}$ converge en loi vers la v.a de loi normale centrée réduite $\mathcal{N}(0;1)$, et on écrit :*

$$\frac{\overline{X_n}-\mu}{\sigma/\sqrt{2\pi}} \underset{n\longrightarrow\infty}{\underset{\longrightarrow}{\mathcal{L}}} \mathcal{N}(0;1).$$

1.2 Test d'hypothèse

Définition 1.2.1 *Un test est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon [5].*

Définition 1.2.2 *Un test d'hypothèse est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires , la validité d'hypothèse relatives à une ou plusieurs population [8] .*

1.2.1 Hypothèse nulle – hypothèse alternative

Définition 1.2.3 ***(Hypothèse nulle)*** *On note H_0 est l'hypothèse que l'on désire contrôler,elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la distribution observée n'est pas significative et est due aux fluctuation d'échantillonnage [8] .*

Remarque 1.2.1 *Cette hypothèse est formulée dans le but d'être rejetée.*

Définition 1.2.4 ***(Hypothèse alternative)*** *on note H_1est la négation de H_0, elle est équivalent à dire "H_0 est $fausse$" [8].*

Remarque 1.2.2 *La décision de rejeter H_0 signifie que H_1 est réalisée ou H_1 est vraie*

–Disons θ_0 pour un paramètre θ de la population, on testera donc

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

Qui test chaque côté de l'égalité(on parlera de test bilatéral), on peut écrire également un autre choix d'hypothèse :

$$H_0 : \theta \geq \theta_0.$$

Et l'hypothèse alternative correspondant sera :

$$H_1 : \theta < \theta_0.$$

Qui test un seul côté de l'égalité(test unilatéral)

Le raisonnement inverse peut être formulé l'hypothèse suivant :

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

Remarque 1.2.3 *$(H_0 : \theta \leq \theta_0$ ou $H_0 : \theta \geq \theta_0)$ parfois noté encore $(H_0 : \theta = \theta_0)$.*

1.2.2 Statistique et niveau de signification

Définition 1.2.5 ***(Une statistique)*** *est une fonction des variables aléatoires représentant l'échantillon dont la valeur numérique obtenue pour l'échantillon considéré permet de distinguer entre H_0 est vraie et H_0 est fausse [7]. Le choix de la statistique dépend de la nature des données [8].*

–Connaissant la loi de probabilité suivie par la statistique S sous l'hypothèse H_0, il est possible d'établir une valeur seuil S_{seuil} de la statistique pour une probabilité donnée appelée le niveau de signification du test [7].

Définition 1.2.6 ***(niveau de signification)*** *Le niveau de signification du test α est la probabilité de dépassement de la valeur observer de la variable de décision sous H_0.*

1.2.3 Région critique et risque d'erreur

Définition 1.2.7 ***(La région critique)*** *on noté W l'ensemble des valeurs de la variable de décision qui conduisent à rejeter H_0 et au profit de H_1 en écrivant que :*

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

Risque d'erreur

Définition 1.2.8 *On appelle risque d'erreur de première espèce la probabilité de rejeter H_0 et d'accepter H_1 alors que H_0 est vraie . Si la valeur de la statistique de test appartient dans la région de rejet alors que l'hypothèse H_0 est vraie. La probabilité de cet évènement est le niveau de signification α ,on écrivant que :*

$$\alpha = P(W/H_0) = P(rejeterH_0/H_0 \ vraie).$$

Exemple 1.2.1 *Si l'on cherche à tester l'hypothèse qu'une pièce de monnaie n'est pas "truquée" , soit X : "nombre de face" obtenue en lançant 100 fois la pièce, nous allons adopter*

la règle de décision suivant :

$$\begin{cases} H_0 : la\ pièce\ n'est\ pas\ truquée. \\ H_1 : la\ pièce\ est\ truquée. \end{cases}$$

H_0 est acceptée si $X \in [40; 60]$;

H_0 est rejetée si $X \notin [40; 60]$, donc soit $X < 40$ ou $X > 60$.Quel est le risque d'erreur de première espèce ?

Si la pièce est truquée, a une probabilité p d'avoir face, et $(1-p)$ d'avoir pile .Si la pièce n'est pas truquée on a $p = 1 - p = \frac{1}{2}$.Donc les hypothèses sont :

$$\begin{cases} H_0 : p = \frac{1}{2}. \\ H_1 : p \neq \frac{1}{2}. \end{cases}$$

Soit X :"nombre de faces obtenues" suit une loi Binomiale telle que $X \sim \mathcal{B}(100, \frac{1}{2})$

$$P(X = k) = \begin{pmatrix} k \\ n \end{pmatrix} p^k (1-p)^{n-k} \quad ; \quad k = 1, 2, 3, ..., n.$$

où : $n = 100$: Nombre de lancés ;

p : Probabilité du succés " obtenir face" , $p = \frac{1}{2}$ la pièce n'est pas truquée ; On calcule α :

$$\alpha = P(\text{rejeter } H_0 / H_0 \text{ est vraie}) = P(X \notin [40, 60] / \text{la pièce n'est pas truquée}).$$

$$\begin{aligned} 1 - P(60 \leq X \leq 84) &= 1 - [P(X \leq 60) - P(X \leq 40)] \\ &= 1 - \left[\sum_{k=1}^{60} P(X = k) - \sum_{k=1}^{40} P(X = k) \right] \\ &= 1 - \left[\sum_{k=1}^{60} \begin{pmatrix} k \\ 100 \end{pmatrix} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{100-k} - \sum_{k=1}^{40} \begin{pmatrix} k \\ 100 \end{pmatrix} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{100-k} \right] \\ &= 1 - \sum_{k=40}^{60} \begin{pmatrix} k \\ 100 \end{pmatrix} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{100-k} = 1 - \sum_{k=40}^{60} \begin{pmatrix} k \\ 100 \end{pmatrix} \left(\frac{1}{2}\right)^{100}. \end{aligned}$$

Méthode de calcul : *On a $X \sim \mathcal{B}(100, \frac{1}{2})$, alors $E(X) = np = 100 \times \frac{1}{2} = 50$;*

et $Var(X) = np(1-p) = 100 \times \frac{1}{2} \times \frac{1}{2} = 25$. On centre et on réduit X, c'est à dire que l'on pose :

$$Z = \frac{X - E(X)}{\sqrt{Var(X)}}$$

donc

$$\begin{aligned}
1-\alpha &= P(X \in [40,60]) \\
&= P(40 \leq X \leq 60) \\
&= P(\frac{40-50}{5} \leq Z \leq \frac{60-50}{5}) = P(-2 \leq Z \leq 2) \\
&= P(Z \leq 2) - P(Z \leq -2) = 2P(Z \leq 2) - 1 = 2 \times 0,977 - 1 = 0.954
\end{aligned}$$

$1-\alpha = 0,954 \Leftrightarrow \alpha = 1 - 0,954 = 0,046 \backsimeq 0.05$.

alors le risque d'erreur de première espèce est $\alpha = 0,05$.

Définition 1.2.9 *On appelle risque d'erreur de deuxième espèce, notée* β *la probabilité de rejeter* H_1 *et d'accepter* H_0 *alors que* H_1 *est vraie. Si la valeur de la statistique de test n'appartient pas dans la région critique alors que l'hypothèse* H_1 *est vraie .*

$$\beta = P(\overline{W} \diagup H_1) = P(\textit{rejeter } H_1 \diagup H_1 \textit{ vraie}).$$

Exemple 1.2.2 *Si l'on reprend l'exemple précédant de la pièce de monnaie ; Soit* X *: "nombre de face" obtenue en lançant 100 fois la pièce et que l'on suppose la probabilité d'obtenir face est de* $0,6$ *pour une pièce truquée. En adoptant toujours la même règle de décision :*

$$\begin{cases} H_0 : \textit{la pièce n'est pas truquée.} \\ H_1 : \textit{la pièce est truquée.} \end{cases}$$

-l'hypothèse H_0 *est acceptée si* $X \in [40;60]$;

-l'hypothèse H_0 *est rejetée si* $X \notin [40;60]$ *donc soit* $X < 40$ *ou* $X > 60$.*Soit* $X \sim \mathcal{B}(100;0,6)$ *où* $P = 0,6$: *probabilité du succès "obtenir face", si la pièce est truquée .*

La loi de probabilité de X est la loi Binomiale $B(n,p)$ telque :

$$E(X) = nP = 100 \times 0,6 = 60; Var(X) = nP(1-P) = 100 \times 0,6 \times 0,4 = 24,$$

Le risque d'erreur de deuxième espèce β est :

$$\beta = P(Accepter\ H_0 / H_1\ est\ vraie) = P(40 \leq X \leq 60).$$

on transfére la v.a X vers la loi $N(0,1)$ (1.2)et on obtient : :

$$\begin{aligned}
\beta &= P(-4,08 \leq Z \leq 0) \\
&= P(Z \leq 0) - P(Z \leq -4,08) \\
&= P(Z \leq 0)) - (1 - P(Z \leq 4,08)) \\
&= 0,5 - (1 - 0,9999) \backsimeq 0,50.
\end{aligned}$$

donc le risque d'erreur de deuxième espèce est $\beta = 0,50$;

On a 50% de chance d'accepter l'hypothèse H_0 " la pièce n'est pas truquée " alors qu'elle est truquée H_1 est vraie.

1.2.4 Règle de décision

Soit H_0 et H_1 ces deux hypothèses, dont une et une seul est vraie. La décision aboutira à choisir H_0 ou H_1, il y a donc quatre cas possibles sont résumées dans le tableau suivant :

Décision$\searrow$Vérité	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

TAB. 1.1 – Tableau présente les cas de régle de décision.

La région d'acceptation est $\overline{W}$ la négation de W :

$$P(\overline{W}/H_0) = 1 - \alpha \text{ et } P(W/H_1) = 1 - \beta \ .$$

Définition 1.2.10 *On appelle puissance d'un test, la probabilité de rejeter H_0 et d'accepter H_1 alors que H_1 est vraie. Sa valeur est $1 - \beta$.*

1.2.5 P-valeur

Définition 1.2.11 ***(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. Aussi pour un seuil de significativité α donnée, on compare P et α, a fin d'accepter, ou rejeter H_0.*

- Si $P \leq \alpha$ on rejet l'hypothèse H_0.
- Si $P > \alpha$ on accepte l'hypothèse H_0.

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

1.3 Test paramétrique

Définition 1.3.1 *Un test paramétrique est un test de contrôler certaine hypothèse relative à un ou plusieurs paramètre 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. Soit $(X_1, X_2, ..., X_n)$ un échantillon issu d'une variable aléatoire de la loi $P(\theta \in \Theta)$. On considéré Θ_0 et Θ_1 avec*

$$\Theta_0 \cup \Theta_1 = \Theta \quad et \quad \Theta_0 \cap \Theta_1 = \varnothing.$$

Donc on a deux hypothèse a tester :

$$\begin{cases} H_0 : \theta \in \Theta_0. \\ H_1 : \theta \in \Theta_1. \end{cases}$$

1.3.1 Test de conformité

Définition 1.3.2 *les tests de conformité sont destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée . Pour une population nous examinons les tests de conformité portant sur un seul paramètre : une moyenne, une variance, une proportion [8].*

– Soit $X_1, ..., X_n$ un échantillon aléatoire de taille n d'une v.a $X \sim \mathcal{N}(\mu; \sigma^2)$ de moyen μ et de variance σ^2, pour un niveau critique α fixée, les tests de conformité sont résumés dans le tableau suivant, [1], selon différents cas :

L'hypothése	**Cas possible**	**La Statistique**	**Région critiques**
$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	σ^2 connue	$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert Z_0 \rvert > z_{\alpha/2}$
	σ^2 inconnue	$T_0 = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert T_0 \rvert > t_{\alpha/2;n-1}$
	σ^2 inconnue, $n \geq 30$	$Z_0 = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert Z_0 \rvert > z_{\alpha/2}$
$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$	μ et σ^2 sont inconnues	$\mathcal{X}_0^2 = (n-1)\frac{S^2}{\sigma_0^2}$	$\mathcal{X}_0^2 \notin \left[\mathcal{X}^2_{1-\alpha/2;n-1}\mathcal{X}^2_{\alpha/2;n-1}\right]$
	$n \geq 30$	$Z_0 = \frac{S - \sigma_0}{\sigma_0/\sqrt{2n}}$	$\lvert Z_0 \rvert > z_{\alpha/2}$
$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$	$X \sim \mathcal{B}(n,p)$ et $n \geq 30$	$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\lvert Z_0 \rvert > z_{\alpha/2}$

TAB. 1.2 – Résumé sur le test de conformité.

Remarque 1.3.1

1. *La Statistique Z_0 suit la loi normale $N(0,1)$;*
2. *La Statistique T_0 suit la loi de Student.de $n-1$ddl et de probabilité $\alpha/2$;*
3. *La Statistique $\mathcal{X}_0^2$ suit la loi de Khi-deux de $n-1$ddl et de probabilité $1-\alpha/2$;*

4. La variance corrigée S égale à

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \bar{X}^2}. \tag{1.3}$$

1.3.2 Test d'homogénéité

Définition 1.3.3 *Les tests d'homogénéité destinés a comparer deux populations à l'aide d'un nombre équivalent d"échantillons. Dans ce cas la loi théorique du paramètre étudié $(p; \mu; \sigma)$ est inconnue au niveau des population étudiées.*

Soit X_1, X_2 deux variables (ou populations) de moyennes μ_1, μ_2 et de variances σ_1^2, σ_2^2. Soit $X_{1,1}, ..., X_{1,n}$; $X_{2,1}, ..., X_{2,n}$ deux échantillons indépendants provenant de X_1 et X_2 respectivement, de moyennes $\overline{X_1}$, $\overline{X_2}$, et de variances S_1, S_2 (1.3). Soit $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$, $i = \overline{1, n}$.[1].

L'hypothése	**Cas possible**	**La Statistique**	**R.C**
$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$	σ_1^2, σ_2^2 connues	$Z_0 = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\lvert Z_0 \rvert > z_{\alpha/2}$
	σ_1^2, σ_2^2 inconnues$(\sigma_1^2 = \sigma_2^2)$	$T_0 = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$	$\lvert T_0 \rvert > t_{\alpha/2; n_1+n_2-2}$
	σ_1^2, σ_2^2 inconnues$(\sigma_1^2 \neq \sigma_2^2)$	$T_0 = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$\lvert T_0 \rvert > t_{\alpha/2; \mathcal{V}}$
$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$	σ_1^2, σ_2^2 inconnues	$F_0 = \frac{S_1^2}{S_2^2}$	$F_0 < \mathcal{F}_{1-\alpha/2; n_1-1; n_2-1}$
	$n_1, n_2 \geq 30$	$Z_0 = \frac{S_1 - S_2}{S_p\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}}$	$\lvert Z_0 \rvert > z_{\alpha/2}$
$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$	$X_i \backsim Bernoulli(p_i)$, $i = 1, 2$ n_1, n_2 trés grands	$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$	$\lvert Z_0 \rvert > z_{\alpha/2}$

TAB. 1.3 – Résumé sur le test d'homogénéité.

Remarque 1.3.2 *1. La statistique Z_0 suit la loi normale $N(0,1)$;*

2. La statistique T_0 suit la loi de Student de n_1+n_2-2 dll et de probabilité $\alpha/2$;

3. La statistique F_0 suit la loi de Fisher de $(n_1-1,\, n_2-1)$ dll et de probabilité $1-\alpha/2$;

4. L'estimateur S_p^2 égale à :

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

Exemple 1.3.1 *Une sondage mené auprès de 600 répondants révèle que 210 de ceux-ci ont l'intension de voter en faveur d'un candidat A aux prochaines élections.*

Peut-on conclure que les intentions de vote pour le candidat A sont supérieures à 25% au niveau critique $\alpha = 0.05$? préciser les hypothèses H_0, H_1.

Soit p la proportion réelle de personnes qui voteront pour le candidat A. On teste

$$\begin{cases} H_0 : p = 0,25 \\ H_1 : p \neq 0,25 \end{cases} \text{, à un seuil critique de } 5\%.$$

$$\hat{P} = \frac{210}{600} = 0,35 \ , \ Z_0 = \frac{0,35-0,25}{\sqrt{\frac{0,25\times(1-0,25)}{600}}} = 5,656 \ , \ z_{0.05/2} = z_{0,025} = 1,96.$$

Puisque $|Z_0| = 0,656 > z_{0,025} = 1,96$, on rejette H_0, oui, on peut conclure que les intentions de vote en faveur du candidat A sont supérieures à 25% .

1.4 Test non paramétrique

Définition 1.4.1 *Un test non paramétrique est un test qui vérifier si la distribution observée d'un échantillon peut être considéré comme extrait d'une population donnée.*

Définition 1.4.2 *Un test d'adéquation permet de statuer sur la compatibilité d'une distribution observée avec une distribution théorique associée à une loi de probabilité.Il s'agit*

de modélisation. nous résumons une information brute , une série d'observations, à l'aide d'une fonction analytique paramétrée. L'estimation des paramètres est souvent un préalable au test de conformité [11] .

1.4.1 Test d'adéquation de Khi-deux

Soit X une variable aléatoire de loi P (le plus souvent inconnue). On souhaite tester l'ajustement de cette loi à une loi connue P_0 (Poisson, Exponentielle, normale,..., etc) retenue comme étant un modèle convenable [9] .

On teste donc les hypothèses :

$$\begin{cases} H_0 : P(x) = P_0(x). \\ H_1 : P(x) \neq P_0(x). \end{cases}$$

Soit X une variable aléatoire discrète ou discrétisée, c'est a dire divisée en k classes de probabilité

$P_1, ..., P_k$ soit un échantillon de cette variable fournissant les effectifs aléatoire $N_1, ..., N_k$ dans chacune de ces classes on a $E(N_i) = nP_i$, ou $i = \overline{1,k}$. Ainsi $\sum_{i=1}^{i=k} N_i = n$. Pour chaque classe, l'effectif théorique est défini :

$$C_i = n.P(X \in Classe_i / X \sim P_0).$$

et la statistique D^2 du test définie comme suit :

$$D^2 = \sum_{i=1}^{i=k} \frac{(N_i - C_i)^2}{C_i}.$$

On rejette H_0 si D^2 constaté supérieure à la valeur théorique $\mathcal{X}^2_\alpha$ lue dans la table du $Khi - deux$ $(D^2 > \mathcal{X}^2_\alpha)$ [6] à $\mathcal{V} = k - 1 - r$ degrés de liberté où r est le nombre de

paramètre de la loi P_0 qu'il a fallu estimer. (Exemple : $r = 0$ si la loi est connue ou imposée, $r = 1$ pour une loi de poisson, $r = 2$ pour une loi normale).

Exemple 1.4.1 *Un pisciculteur possède un bassin qui contient trois variétés de truites : communes, saumonées et arc-en-ciel. Il voudrait savoir s'il peut considérer que son bassin contient autant de truites de chaque variété. Pour cela, il effectue, au hasard* 399 *prélèvements avec remise et obtient les résultats suivants :*

Variétés	commune	saumonée	arc-en-ciel
Effectifs	145	118	136

TAB. 1.4 – Resultats de chaque variété de truite.

On cherche à savoir s'il y a équirépartition des truites entre chaque espèce c'est-à-dire on suppose de P_0 *est la loi uniforme,* $n = 399$*, une probabilité de* $1/3$ *pour chaque classe* (*soit* $C_i = 399 \times \frac{1}{3}$)

Variétés	commune	saumonée	arc-en-ciel
Effectifs O_i	145	118	136
Effectifs C_i	133	133	133

TAB. 1.5 – Effectifs théoriques de chaque variété de truite.

On obtient :

$$D^2 = \frac{(145-133)^2}{133} + \frac{(118-133)^2}{133} + \frac{(136-133)^2}{133} \backsimeq 2.84 \ .$$

La valeur théorique lue dans la table du $\mathcal{X}^2$ *au risque de* 5% *avec* $\mathcal{V} = 3-1-0 = 2$ *degrés de liberté vaut* $5,99$ *.On ne On ne peut rejeter l'hypothèse que son bassin contient autant de truites de chaque variété car* $(D^2 < \mathcal{X}^2)$.

Chapitre 2

Tests de normalité

Les Tests de normalité sont des cas spéciaux des test d'adéquation. Ils sont destinée à examiner la compatibilité d'une distribution empirique avec la loi normale ; c'est à dire si des données réelles suivent une loi normale ou non. Dans ce chapitre, nous allons donner les techniques statistiques les plus populaires pour faire le test de normalité, les méthodes théoriques comme (test de Kolmogorov-Smirnov, test de lilliefors ,...,etc), et ils se concentrent sur les principales formules , aussi les méthodes graphiques comme(Q-Q plot, Histogramme,...,etc)

Définition 2.0.3 *Le test de normalité est un test non paramétrique des hypothèses* :

$$\begin{cases} H_0 : \textit{Loi de } X \textit{ est une normale.} \\ \textit{contre} \\ H_1 : \textit{Loi de } X \textit{ n'est pas une normale.} \end{cases}$$

2.1 Méthodes graphiques

2.1.1 Histogramme de fréquence

L'histogramme de fréquence s'agit de couper automatiquement l'intervalle de définition de la variable en k intervalles de largeur égales, puis de produire une série de barres dont la hauteur est proportionnelle à l'effectif associé à l'intervalle.

Une règle simple pour définir le bon nombre d'intervalles est d'utiliser la règle $k = \log(n)$, où n est la taille de l'échantillon. En l'utilise pour comparer la distribution des données analysée en les représentants sous forme d'histogramme avec une courbe représentant la loi normale.

Exemple 2.1.1 *Pour une population Ω, nous voulons étudier la conformité de la distribution pour chaque variable aléatoire continue $(X_1,\ X_2)$ avec la loi normale. Nous disposons cela de $n_1 = 30$ et $n_2 = 30$ observations suivants :*

$X_1 = (14, 32, 6, 13, 11, 2, 12, 12, 13, 30, 21, 0, 9, 20, 17, 6, 13, 10, 2, 10, 17, 14, 23, 22, 13, 21, 18, 16, 27, 20)$;

$X_2 = (26, 26, 30, 9, 27, 8, 31, 33, 22, 32, 26, 31, 35, 6, 0, 13, 33, 0, 34, 35, 12, 0, 17, 0, 17, 27, 8, 11, 18, 2)$;

Est ce que la distribution des échantillons X_1 et X_2 suit une loi normale ?

On peut représenter les données à l'aide de l'histogramme de fréquence et regarder si elles semblent s'ajuster à une distribution normale.

Code R [2] :

```
X1=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)
X2=c(26,26,30,9,27,8,31,33,22,32,26,31,35,6,0,13,33,0,34,35,12,0,17,0,17,27,8,11,18,2)
par(mfrow=c(1,2))
hist(X1,main="Histogramme de X1",ylab="fréquence", prob=T)
```

curve(dnorm(x, mean(X1), sd(X1)),add=TRUE)

hist(X2,main="Histogramme de X2",ylab="fréquence", prob=T)

curve(dnorm(x, mean(X2), sd(X2)),add=TRUE)

Résultat de la commande :

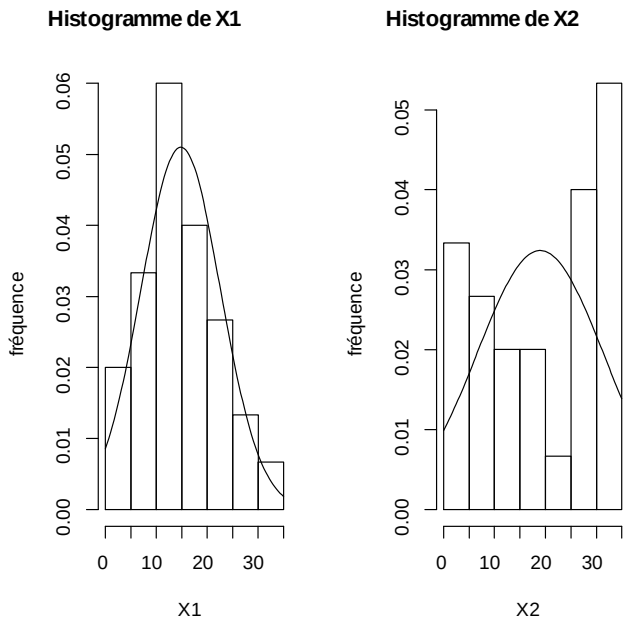

FIG. 2.1 – Histogramme de fréquence avec densité de la loi normale de X1, X2

Dans (figure 2.1) on observe que les données de v.a X_1 sont centrées et semblent s'ajuster à la courbe de la loi normale, et les données de v.a X_2 sont plus dispersées et s'éloigne plus fortement de la loi normale.

2.1.2 Boite à moustache

La boite à moustaches (*box plot*) est un outil graphique très pratique représentant une distribution empirique à l'aide de quelques paramètres de localisation : la médiane(M), le premier quartile(Q_1)et troisième quartile(Q_3).

– La médiane : c'est la valeur "centrale" de la série. On dit qu'elle partage la série en deux moitiés.

– Les quartiles : partagent la série en 4, Il y en a donc :

1. Le première quartile ($1^{ère}$ Q_1) est la plus petite valeur, telle que 25% des données lui soit inférieures ou égales.
2. Le deuxième quartile $\left(2^{ème}\ Q_2\right)$ est la médiane.
3. Le troisième quartile $\left(3^{ème}\ Q_3\right)$ est la plus petite valeur, telle que 75% des données lui soit inférieures ou égales.

Lire la boite à moustache

1. L'intervalle $[Q_1, Q_3]$ est s'appelle l'intervalle interquartile.
2. Le nombre $Q_3 - Q_1$ s'appelle l'écart interquartile.
3. **La moustache inférieure :** valeur de la série immédiatement supérieure à la frontière basse, avec la frontière basse qui est égale à $Q_1 - 1,5 \times (Q_3 - Q_1)$.
4. **La moustache supérieure** : valeur de la série immédiatement inférieure à la frontière haute, avec la frontière haute égale à $Q_3 + 1,5 \times (Q_3 - Q_1)$.

Notation 2.1.1 *Une boite à moustache est dite symétrique lorsque la position de la médiane se situe au milieu de la boite à moustache et qu'il y a symétrie des moustaches.*

Exemple 2.1.2 *On prend l'exemple précédent (2.1.1) ; La boit à moustache représente une distribution empirique à l'aide des quartiles : minimum* (Min)*, maximum* (Max)*, médiane* $(Median)$*,* 1^{er} *quartile* $(1st\ Qu)$*,* $3^{ème}$ *quartile* $(3rd\ Qu)$ ***[2].***

code R :

```
X1=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)
X2=c(26,9,30,33,24,8,33,8,33,10,13,24,35,17,13,0,13,14,13,13,12,0,17,35,17,27,8,23,18,21)
summary(X1)
Min. 1st Qu. Median  Mean   3rd Qu.  Max.
```

0.00 10.25 13.50 14.80 20.00 32.00

summary(X2)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 12.25 17.00 18.23 25.50 35.00

boxplot(X1,X2,names=c("X1","X2"))

Résultat da la commande :

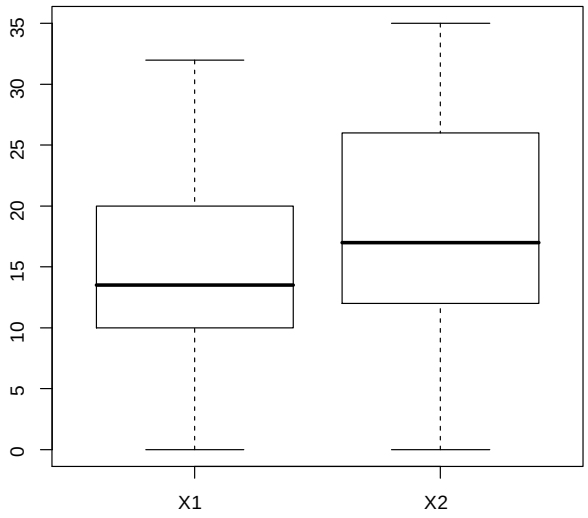

FIG. 2.2 – Boite à moustache de X1, X2

On observe dans (figure 2.2) :

pour une v.a X_1, on a $(M = 13.5,\ Q_1 = 10.25,\ Q_3 = 20,\ Min = 0,\ Max = 32)$ la médiane de X_1 se situe légèrement dans la partie inférieur de la boite à moustache et que le minimum et le maximum sont légèrement asymétrique.

Pour une v.a X_2, on a $(M = 17,\ Q_1 = 12.25,\ Q_3 = 25.50,\ Min = 0,\ Max = 35)$.

Conclusion 2.1.1 *La boite à moustache permet d'observer les valeurs extrême mais également d'avoir une idée sur la symétrie de la distribution. La symétrie d'une distribution*

n'affirme pas la normalité, mais une distribution normale est forcément symétrique.

2.1.3 Q-Q plot- Droite de Henry

Q-Q plot :

Le $Q-Q$ *plot*, quantile-quantile plot, est une technique graphique qui permet de comparer les distributions de deux ensembles des données [**??**] .

Il est un graphique qui permet de tester la conformité entre les quantiles d'une distribution empirique d'une variable et les quantiles d'une distribution théorique données.

–Dans cet champ nous appliquerons au test de conformité à la distribution normale. Il s'agit :

1. de trier les données de manière croissante pour former la série $X_{(i)}$;
2. à chaque valeur $X_{(i)}$, nous associons la fonction de répartition empirique :

$$F_i = \frac{i - 0,375}{n + 0,25}. \tag{2.1}$$

3. nous calculons les quantiles successifs $z_{(i)}$ d'ordre F_i en utilisant l'inverse de la loi normale centrée et réduite.
4. en fin, les données initiales n'étant pas centrées et réduites, nous dé-normalisons les données en appliquant la transformation

$$x_{*(i)} = z_{(i)} \times S + \overline{X}. \tag{2.2}$$

où $\overline{X}$ est l'estimateur de la moyenne (moyenne empirique) et S est l'estimateur de l'écart type σ et sont :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad ; \quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}. \tag{2.3}$$

Si les données sont compatibles avec la loi normale, les points $(x_{(i)},\ x_{*(i)})$ forment une droite, dite droite de Henry, alignés sur la diagonale principale.

Droite de Henry :

La droite de Henry est une méthode pour visualiser les chances qu'a une distribution d'être gaussienne .Elle permet de lire rapidement la moyenne et l'écart type d'une telle distribution [9].

–Le principe est représenté les quantiles théorique en fonction des quantiles observés (Diagramme Q-Q).

Si X est une variable gaussienne de moyenne $\overline{X}$ et si Z est une variable de loi normale centrée et réduite, on a les égalités suivant :

$$P(X < x_i) = P\left(\frac{X - \overline{X}}{\sigma} < \frac{x_i - \overline{X}}{\sigma}\right) = P(Z < y_i) = \Phi(y_i)\ . \tag{2.4}$$

Où Φ est la fonction de répartition de loi normale centré réduite, et $y = \frac{x-\overline{X}}{\sigma}$;

Pour chaque x_i de la variable X , on peut

1. Calculer $P(X < x_i\ ),\ i = 1, ..., n$;
2. À l'aide d'une table de la fonction Φ en déduire y_i

$$y_i = \Phi^{-1}(P(X < x_i)).$$

Si la variable est gaussienne, les points de coordonnées (x_i, y_i) sont a alignées sur la droite d'équation :

$$y = \frac{x - \overline{X}}{\sigma}.$$

On compare donc les valeur des quantiles de loi empirique (x_i) aux quantiles de loi normale centrée et réduite (y_i) [6].

Exemple 2.1.3 *On prend même exemple (2.1.1), [2].*

Code R :

```
X1=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)
X2=c(26,9,30,33,24,8,33,8,33,10,13,24,35,17,13,0,13,14,13,13,12,0,17,35,17,27,8,23,18,21)
par(mfrow=c(1,2))
qqnorm(X1, datax=TRUE)
qqline(X1, datax=TRUE)
qqnorm(X2, datax=TRUE)
qqline(X2, datax=TRUE)
```

Résultat da la commande :

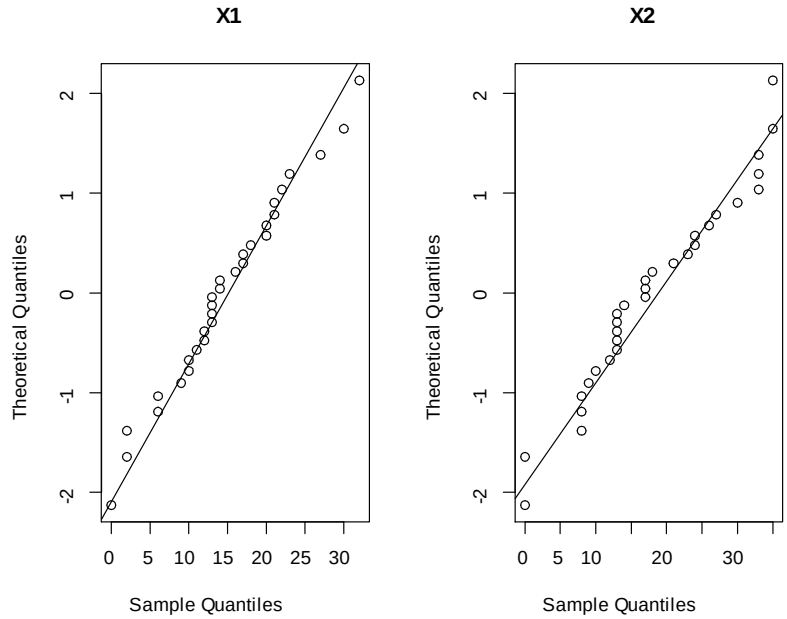

FIG. 2.3 – Q-Q plot et droite de Henry pour X1, X2

Nous obtenons les graphiques nuages des points (figure 2.3), les données de X_1 sont proche de la droite nous constatons que les points sont relativement alignés tandis que les données de X_2 sont plus éloignées, nous observons que écartement significatif, il y a des points

semble démarquer des autres. Les données de X_1 se rapprochent de la droite, alors la distribution empirique est dite normale.

Conclusion 2.1.2 *En utilisant la méthode graphique, nous avons trouvé que la distribution empirique de X_1 proche de la loi normale, et la distribution empirique de X_2 ne suit pas de la loi normale.*

2.2 Méthodes théoriques

2.2.1 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est d'adéquation pour des variables aléatoires continues, l'objectif est d'établir la plausibilité de l'hypothèse selon laquelle l'échantillon a été prélevé dans une population ayant une distribution données [9].

AOn utilise le test de Kolmogorov-Smirnov pour tester la normalité d'un échantillon, alors la loi donnée dans notre cas est la loi normale.

Etant donnés un échantillon $(x_1, ..., x_n)$ indépendantes identiquement distribuées de taille n d'une variable aléatoire X dont la fonction de répartition F, inconnue et continue. On souhaite tester :

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

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

$$KS = \sup_{x \in \mathbb{R}} |F_{emp}(x) - F_0(x)| = \max_i \max \left\{ \left| F_0\left(x_{(i)}\right) - \frac{i}{n} \right| ; \left| F_0\left(x_{(i)}\right) - \frac{i-1}{n} \right| \right\},$$

où $(x_{(i)})_{i=1}^n$ est l'échantillon ordonné et la fonction de répartition empirique F_{emp} est

la proportion des observation dont la valeur est inférieure à x, elle est définie par :

$$F_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i \leq x)}.$$

et F_0 : la fonction de répartition de la loi normale.

2. **La région critique du test** : au seuil α donné, on accepte l'hypothèse d'égalité de la loi normale si :

$$KS \leq KS_{n,1-\alpha};$$

où la valeur $(KS_{n,1-\alpha})$ étant donné par la table qantile de Komogorov-Smirnov [12].

– Une valeur élevée de KS est une indication que la distribution de l'échantillon s'éloigne sensiblement de la distribution de référence $F(x)$, et qu'il est donc peu probable que H_0 soit correcte. Plus précisément

$$P(\sup_{x \in \mathbb{R}} |F_{emp}(x) - F_0(x)| > \frac{c}{\sqrt{n}}) \underset{n \longrightarrow \infty}{\longrightarrow} \alpha(c) = 2\sum_{r=1}^{+\infty}(-1)^{r-1}\exp(-2r^2c^2),$$

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 usuelles de c en fonction de α sont :

α	$0,20$	$0,10$	$0,05$	$0,02$	$0,01$
c	$1,073$	$1,223$	$1,358$	$1,518$	$1,629$

TAB. 2.1 – Les valeurs de c pour calculer la valeur critique du test.

On rejette H_0 si :

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

Remarque 2.2.1 *La statistique KS de test est basée sur la distance maximale entre la fonction de répartition empirique et F_0.*

Exemple 2.2.1 *On va tester des d'hypothèses suivants, on pose $X = X_1$ selon l'exemple (2.1.1), [10] :*

$$\begin{cases} H_0 = \textit{la distibution de la variable } X \textit{ suit une loi normale.} \\ H_1 = \textit{la distibution de la variable } X \textit{ ne suit pas la loi normale.} \end{cases}$$

Calculs du test :

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

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S}, \ \textit{où} \ \ \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = 14.8 \ \textit{et} \ \ S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} = 7.8186 \tag{2.5}$$

3. *trouver les valeur de F correspondantes avec les valeur calculées à l'étape précédente (2.4).*
4. *calculer la valeur maximale de :*

$$\max_i \left| F_0\left(x_{(i)}\right) - \frac{i}{n} \right| = 0.10742 \ \textit{et} \ \max_i \left| F_0\left(x_{(i)}\right) - \frac{i-1}{n} \right| = 0.07408 \ .$$

alors

$$KS = \sup_{x \in \mathbb{R}} |F_{emp}(x) - F_0(x)| = 0.10742 \ .$$

5. *au seuil $\alpha = 0.05$ et pour $n = 30$ on comparer :*

$$KS = 0.10742 < KS_{n,1-\alpha} = 0,2417.$$

Code R :

```
X=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)
ks.test(X, pnorm, mean=mean(X), sd=sd(X))
```

One-sample Kolmogorov-Smirnov test

data : X

D = 0.10742, p-value = 0.8793

alternative hypothesis : two-sided

Commentaire : On remarque que la $P - value$ est supérieure au niveau α, alors On accepte l'ypothèse H_0. Alors on accepte l'ypothèse H_0, les données sont compatibles avec l'hypothèse de normalité.

2.2.2 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 ($n \leq 50$) [11] .

La statistique du test s'écrit :

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i \left(x_{(n-i+1)} - x_{(i)}\right)\right]^2}{\sum_i \left(x_{(i)} - \overline{X}\right)^2}. \tag{2.6}$$

où

- $x_{(i)}$ correspond à la série des données triées ;
- $\left[\frac{n}{2}\right]$ est la partie entière du rapport $\frac{n}{2}$;
- les a_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 de la loi normale et les quantile empirique obtenues à partir des données, plus la compatibilité avec la loi normale est crédible [11].

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

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

Les valeurs seuils W_{crit} pour différents risque α et effectifs n sont lues dans la table de Shapiro-Wilk [**??**] .

- Si la $P-value$ est inférieure à un niveau α choisi alors l'hypothèse nulle est rejetée c'est à dire improbable d'obtenir de telle données en supposant qu'elles soient normalement distribuées.
- Si $P-value$ est supérieure au niveau α choisi alors on ne doit pas rejeter l'hypothèse nulle.La valeur de la $p-valeur$ alors obtenue ne présuppose en rien de la nature de la distribution des données.

  Exemple 2.2.2 *Les calculs s'agencent de la manière suivante avec l'exemple (2.1.1), [10] :*

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

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

  3. *Lire dans la table pour $n = 30$ et $i = \overline{1, 15}$ les valeur de coefficient a_i [3];*
  4. *Former le numérateur de W :*

$$d^2 = \left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i \left(x_{(n-i+1)} - x_{(i)}\right)\right]^2 = 1735.075;$$

  5. *Former le dénominateur de W :*

$$S^2 = \sum_i \left(x_{(i)} - \overline{X}\right)^2 = 1772.8;$$

6. En déduire la valeur de W :

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i \left(x_{(n-i+1)} - x_{(i)}\right)\right]^2}{\sum_i \left(x_{(i)} - \overline{X}\right)^2} = 0.97872;$$

7. Pour une risque α, le seuil critique lue dans la table [3]:

$$W_{crit} = 0,927.$$

8. Comparer entre W et W_{crit} :

$$W > W_{crit}.$$

Code R :

X=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)

shapiro.test(X)

Shapiro-Wilk normality test

data : X

W = 0.97872, p-value = 0.7906

Commentaire : On remarque que $P - value$ est supérieure au niveau α, ce qui confirme la validité de l'hypothèse H_0. Alors on accepte H_0, c'est à dire les données suivent une distribution normale.

2.2.3 Test de lilliefors

Le test de Lilliefors est une variante du test de Kolmogorov-Smirnov où les paramètres de la loi (μ *et* σ) sont estimés à partir des données [11].

La statistique du test : s'écrit :

$$D = \max_{i=1,\dots,n} \left(F_i - \frac{i-1}{n}, \frac{i}{n} - F_i \right).$$

où $F_i = F(x_i)$ est la fréquence théorique de la loi de répartition normale centrée et réduite associée à la valeur standardisée $z_{(i)} = \frac{x_{(i)} - \overline{x}}{S}$.

Valeurs critiques : La table des valeurs critique D_{crit} pour les petites valeur de n et différentes valeur 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	$Valeur\ critique\ D_{crit}$
$0,10$	$\frac{0,805}{\sqrt{n}}$
$0,05$	$\frac{0,886}{\sqrt{n}}$
$0,01$	$\frac{1,031}{\sqrt{n}}$

TAB. 2.2 – Les valeur critique du test de Lilliefors en fonction de n.

Région critique du test : pour la statistique D elle est définie par :

$$R.C : D > D_{crit}.$$

Exemple 2.2.3 *On résume les étapes de ce test comme suit et on l"applique sur l'exemple (2.1.1) :*

1. *Les données sont triées pour former la série* $x_{(i)}$;
2. *Estimer les paramètres* $\overline{X}$ *et* S (2.5)
3. *Calculer alors les données centrés réduites :*

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S};$$

4. *Utiliser la fonction de répartition de la loi normale centrée et réduite pour obtenir les fréquence théorique* F_i, $i = \overline{1,n}$;

5. *Que nous opposons aux fréquence empiriques pour obtenir la statistique* D *du test, en calculant*

$$D^- = \max_{i=\overline{1,n}}(F_i - \frac{i-1}{n}) = 0.069634 \quad puis \quad D^+ = \max_{i=\overline{1,n}}(\frac{i}{n} - F_i) = 0.107415;$$

et enfin

$$D = \max_{i=\overline{1,n}}(D^-,\ D^+) = 0.10742;$$

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

$$D < D_{crit};$$

Code R :

X=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)

lillie.test(X)

Lilliefors (Kolmogorov-Smirnov) normality test

data : X

D = 0.10742, p-value = 0.5064

Commentaire : On remarque que la $p - value$ est supérieure à une risque α,donc on accepté l'hypothèse H_0. Alors les données sont compatibles avec l'hypothèse de normalité.

2.2.4 Test d'Anderson-Darling

Le test d'Anderson-Darling est une autre variante du test de Kolmogorov-Smirnov, à la différence qu'elle donne plus d'importance aux queues de distribution. De ce point de vue, elle est plus indiquée dans la phase d'évaluation des données précédant la mise en oeuvre

d'un test paramétrique (comparaison de moyenne, de variante,...) que
le test de Lilliefors [11].

Autre particularité, ses valeurs critiques sont tabluées différemment selon la loi théorique de référence, un coefficient multiplicatif correctif dépendant de la taille d'échantillon n peut être introduit.

La statistique du test : S'écrit :

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

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

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S}.$$

Valeurs critiques : 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. 2.3 – Les valeusr critiques du test d'Aderson-Darling.

Remarque 2.2.2 *Les valeurs critiques A_{crit} ont été produits par simulation et ne dépendent pas de l'effectif de l'échantillon.*

Région critique du test : L'hypothèse de normalité est rejetée si :

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

La $P-value$ est calculer à partir de la statistique A_m :

$$A_m = A\left(1 + \frac{0,75}{n} + \frac{2,25}{n^2}\right).$$

Puis on utilisé la règle suivant pour déduire la $P-value$:

A_m	$P-value$
$A_m < 0.2$	$1-\exp(-13.436+101.14*A_m-223.73*(A_m)^2)$
$0.2 \leq A_m < 0.34$	$1-\exp(-8.318+42.796*A_m-59.938*(A_m)^2)$
$0.34 \leq A_m < 0.6$	$\exp(0.9177-4.279*A_m-1.38*(A_m)^2)$
$0.66 \leq A_m$	$\exp(1.2937-5.709*A_m+0.0186*(A_m)^2)$

TAB. 2.4 – Les règles de P-value selon de la statistique Am.

Exemple 2.2.4 *Les étapes du test pour l'echantillon X (2.1.1)sont présentés comme suit :*

1. *Les données sont triées pour former la série* $x_{(i)}$ *;*
2. *Estimer les paramètres* $\overline{X}$ *et* S (2.5) ;
3. *calculer les données centrées et réduites :*

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S};$$

4. *Utilisons la fonction de répartition de la loi normale centrée et réduite pour obtenir les fréquences théorique* F_i*, et calculer* $\ln(F_i)$;
5. *former* F_{n-i+1} *puis en déduire* $\ln(1-F_{n-i+1})$;
6. *calculer la somme :*

$$s = \sum_{i=1}^{n}(2i-1)\left[\ln(F_i) + \ln(1-F_{n-i+1})\right] = -907.3936,$$

puis calculer la statistique

$$A = -n - \frac{1}{n}s = 0.24645;$$

7. *Comparer au seuil critique* $A_{crit} = 0.752$ *à une risque* $\alpha = 0.05$ *:*

$$A < A_{crit}.$$

8. *calculer* $P - value$ *pour* $n = 30$:

$$\begin{cases} A_m = A\left(1 + \frac{0,75}{n} + \frac{2,25}{n^2}\right) = 0.25323, \\ 0.2 \leq A_m < 0.34, \text{ alors, } P - value = 1 - \exp(-8.318 + 42.796 * A_m - 59.938 * (A_m)^2) = 0.734 \end{cases}$$

Code R :

X=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)

ad.test(X)

Anderson-Darling normality test

data : X

A = 0.24645, p-value = 0.734.

Commentaire : On remarque que la $p - value$ est supérieure à une risque α.Alors on accepte l'hypothèse H_0, la distribution de la variable X suit une loi normale.

2.2.5 Test de Cramer-Von Mises

Le test de Cramer-Von Mises est un test statistique utilisé pour évaluer la qualité de l'adéquation d'une fonction de répartition F comparée à une fonction de répartition empirique F_{emp}, ce test est également une alternative au test de Kolmogorov-Smirnov.

La statistique du test : est définie par :

$$W_n^2 = n \int_{-\infty}^{+\infty} |F_{emp}(x) - F_0(x)|^2 \, dF_0(x),$$
$$W_n^2 = \sum_{i=1}^{n} \left(F_0(x_{(i)}\right) - \frac{2i-1}{2n})^2 + \frac{1}{12n}.$$

Région critique du test : On rejette H_0 si

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

Pour un niveau α données. La valeur de W_{crit} calculée à partir la table de Cramer-Von Mises [4].

Exemple 2.2.5 *pour un échantillon X et l'exemple (2.1.1), les étapes de calcules sont les suivants :*

1. *trier des données en ordre croissant $(x_i)_{i=1}^n$;*
2. *estimer les paramètres $\overline{X}$ et S (2.5) ;*
3. *calculer les données centrées et réduites :*

$$z_{(i)} = \frac{x_{(i)} - \overline{X}}{S}, \ i = \overline{1,n};$$

4. *utilisons la fonction de répartition de la loi normale centrée et réduite pour exrtraire les fréquences théorique F_i;*
5. *calculer la statistique :*

$$W_n^2 = \sum_{i=1}^{n} \left(F_0(x_{(i)}\right) - \frac{2i-1}{2n})^2 + \frac{1}{12n} = 0.039666.$$

6. comparer au seuil $\alpha = 0.05$, *et pour* $n = 30$, $W_{crit} = 0.218$ *et la statistique* W_n^2 :

$$W_n^2 < W_{crit}.$$

Code R :

X=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)

cvm.test(X)

Cramer-von Mises normality test

data : X

W = 0.039666, p-value = 0.6759

Commentaire : On remarque que la $p - value$ est supérieure à une risque α. Alors on ne rejette pas l'hypothèse H_0, la distribution de la variable X est compatible avec la loi normale.

2.2.6 Test de Jarque-Bera

La loi normale est caractérisée par un coefficient d'asymétrie et un coefficient d'aplatissement nulles, il parait naturel de calculer ces indicateur pour se donner une idée, ne serait-ce que très approximation du rapprochement possible de la distribution empirique avec une gaussienne [11]. Avant de présenter le principe de ce test on va définir le coefficient d'asymétrie et le coefficient d'aplatissement.

Coefficient d'asymétrie

Le coefficient d'asymétrie (skewness en anglais) est une mesure de l'asymétrie de la distribution d'une v.a réelle. c'est le premier des paramètres de forme.

Soit X une v.a réelle de moyenne μ et d'écart σ, on définit son coefficient d'asymétrie

comme le moment d'ordre trois de la variable centrée réduite :

$$\beta_1 = \frac{E(X-\mu)^3}{\left(E(X-\mu)^2\right)^{3/2}} = \frac{E(X-E(X))^3}{\left(E(X-E(X))^2\right)^{3/2}} = \frac{\mu^3}{\sigma^3}.$$

Forme de la distribution :

- Un coefficient nulle indique une distribution symétrique.
- Un coefficient négatif indique une distribution décalée à droite de la médiane, et donc une queue de distribution étalée vers la gauche.
- Un coefficient positif indique une distribution décalée à gauche de la médiane, et donc une queue de distribution étalée vers la droite.

Estimation de l'asymétrie :

L'estimation nous biaisé pour la loi normale est :

$$\hat{\beta}_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{X}}{S}\right)^3.$$

Coefficient d'aplatissement :

Le coefficient d'aplatissement (Kurtosis) est une mesure directe de l'acuité et une mesure indirecte de l'aplatissement de la distribution d'une réelle, c'est le deuxième des paramètres de forme.

Soit X une v.a réelle d'espérance μ et d'écart type σ, on définit son coefficient d'aplatissement non normalisé comme le moment d'ordre quatre de la variable centrée réduite :

$$\beta_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{E(X-E(X))^4}{(E(X-E(X))^2)^2} = \frac{\mu^4}{\sigma^4}.$$

Estimation de l'aplatissement :

L'estimation de de coefficient d'aplatissement est définie par :

$$\hat{\beta}_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{X}}{S} \right)^4 - 3\frac{(n-1)^2}{(n-2)\,(n-3)}.$$

Remarque 2.2.3 *Le coefficient d'aplatissement n'est pas normalisé* $\beta_2 = 3$.

La loi conjointe de ces estimateurs est normale bivariée :

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} \overset{loi}{\rightsquigarrow} \mathcal{N}\left[\begin{pmatrix}0 \\ 3\end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix}\right]$$

La matrice de covariance présentée ici est une expression simplifiée valable pour les grandes valeurs de n. Il est possible de produire des expressions plus précises, affichées par les logiciels de statistique.

Nous notons $(Cov(\hat{\beta}_1, \hat{\beta}_2) = 0)$ c'est la covariance de $\hat{\beta}_1$ et $\hat{\beta}_2$.

Principe du test :

Le test de normalité de Jarque-Bera est également fondé sur les coefficients d'asymétrie et d'aplatissement. Il évalue les écarts simultanés de ces coefficients avec les valeurs de référence de la loi normale. L'hypothèse de ce test est le suivant :

$$\begin{cases} H_0 : \beta_1 = 0 \ et \ \beta_2 = 3 \\ H_1 : \beta_1 \neq 0 \ et \ \beta_2 \neq 3 \end{cases}$$

– La statistique de test Jarque-Bera, noté T, est définie par :

$$T = n\left(\frac{\hat{\beta}_1^2}{6} + \frac{(\hat{\beta}_2 - 3)^2}{24}\right) = n\left(\left(\frac{\hat{\beta}_1}{\sqrt{6}}\right)^2 + \left(\frac{\hat{\beta}_2 - 3}{\sqrt{24}}\right)^2\right),$$

telle que la distribution asymptotique de la statistique T est la loi de Khi-deux $(\mathcal{X}_2)$ de

degrés de liberté ($ddl = 2$) .

– La région critique pour un risque α de ce test est définie par :

$$R.C. : T > \mathcal{X}^2_{1-\alpha}(2).$$

où $\mathcal{X}^2_{1-\alpha}(2)$ est une valeur théorique lu à partir la table de $\mathcal{X}^2$ (6).

Exemple 2.2.6 *Les étapes du test pour une v.a X :*

1. *nous calculons la moyenne empirique :*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = 12.85714;$$

2. *nous formons*

$$d = x - \overline{X};$$

puis d^2, d^3, et d^4;

3. *calculons successivement les estimateurs β_1 et β_2 :*

$$\hat{\beta}_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{S}\right)^3 = 0.206629;$$
$$\hat{\beta}_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{S}\right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)} = 2.778345.$$

4. *calculer la statistique T :*

$$T = n\left(\frac{\hat{\beta}_1^2}{6} + \frac{(\hat{\beta}_2 - 3)^2}{24}\right) = 0.27489.$$

5. *comparer entre la statistique T et le seuil critique de $\mathcal{X}^2_{0.05}(2) = 5.99$ (6):*

$$T < \mathcal{X}^2_{0.05}(2).$$

Code R :

```
X=c(14,32,6,13,11,2,12,12,13,30,21,0,9,20,17,6,13,10,2,10,17,14,23,22,13,21,18,16,27,20)
library(moments)
skewness(X)
[1] 0.206629
kurtosis(X)
[1] 2.778345
library("tseries")
jarque.bera.test(X)
Jarque Bera Test
data : X
X-squared = 0.27489, df = 2, p-value = 0.8716
```

Commentaire : On remarque que la valeur de $p - value = 0.3908$, il est supérieure à niveau de signification. Alors on accepte l'hypothèse nulle H_0, la distribution observée est compatible avec une distribution théorique normale.

Conclusion

Le but de notre mémoire est la représentation des méthodes destinées à évaluer la compatibilité d'une distribution empirique avec la loi normale. Il existe deux méthodes pour vérifier la normalité. La première méthode est la méthode graphique, elle illustre la forme de la distribution de l'échantillon réel, telle que la densité de la loi normale selon le graph de l'histogramme qui est caractérisé par la symétrie ainsi que la boite à moustache mais elle n'est pas seule car il existe d'autres distributions comme la loi de student et de Cauchy, alors on ne peut pas se fier entièrement à la méthode graphique pour vérifier la normalité. Nous nous sommes donc appuyés sur des tests d'hypothèse. La deuxième méthode est la méthode théorique, il existe également un grand nombre de tests de normalité :

- Tests basés sur la fonction de répartition empirique, test de Kolmogorov- Smirnov et son adaptation le test de Lilliefors, test d'Anderson-Darling et test de Cramer-Von Mises ;
  - Test basés sur les moments comme le test de Jarque-Bera, test de D'Agostino ;
  - Ou encore test de Shapiro-Wilk.

On résume les étapes des tests de normalité dans les points suivant :

1. Trier des données pour former la série d'échantillon X ;
2. fixer une valeur de seuil critique α; puis poser les hypothèses :

$$\begin{cases} H_0 = la\ distribution\ de\ la\ variable\ X\ suit\ une\ loi\ normale. \\ H_1 = la\ distribution\ de\ la\ variable\ X\ ne\ suit\ pas\ la\ loi\ normale. \end{cases}$$

3. déterminer un test parmi les tests de normalité, ensuite calculer la statistique de ce test choisi ;
4. comparer la statistique avec les valeurs critiques à partir des tables correspond à chaque statistique, par rapport à la région de rejet l'hypothèse de normalité H_0 . La $p - value$ est souvent utilisée pour la comparer avec le seuil α, si $p - value > \alpha$, on accepte H_0, si non on le rejette.

Bibliographie

[1] AKAKPO, N. (7 SEPTEMBRE 2017). Tests statistiques. MASTER 1 MATHÉMATIQUES ET APPLICATIONS UNIVERSITÉ PIERRE ET MARIE CURIE.

[2] Chesneau, C. (2016). Introduction aux graphiques avec R.

[3] Christophe, C. Tables de valeurs, https ://chesneau.users.lmno.cnrs.fr/tables-valeurs.pdf.

[4] Critical Values for Cramér-von Mises Test. (2009), (file :///D :/Cramer-von%20Mises.pdf).

[5] Gilbert, S. (2006). Probabilités, analyse des données et satistique. Editions Technip.

[6] GOVAERTS, B. (2016). Tables de probabilités.

[7] Jean-Jacques, R. (2013). STATISTIQUE. Préparation à l'Agrégation Bordeaux 1.

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

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

[10] PONGER, L. (6 mars 2012). Les tests statistiques élémentaires avec R.

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

[12] TABLES DE PROBABILITES ET STATISTIQUE, (20 décembre 2013), http ://wwwmathlabo.univ-poitiers.fr/~phan/downloads/enseignement/tables-usuelles.

Annexe A : Logiciel *R*

2.3 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'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

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

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

Annexe B : Abréviations et Notations

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

$\overline{X_n}$	:	Moyenne empirique.
$\mathcal{L}$	:	Convergence en loi.
$v.a$	:	Variable aléatoire.
ddl	:	Degré de liberté.
S	:	Variance empirique.
$Var(X)$	:	Variance théorique.
$E(X)$	:	Moyenne théorique.
$R.C$	:	Région critique.
H_0	:	Hypothèse nulle.
H_1	:	Hypothèse alternative.
F_{emp}	:	Fonction de répartition empirique.
Ω	:	Ensemple de population

Résumé :

La loi normale joue un role très important en statistique également en probabilité à cause que la plupart des données en réalité suivent la distribution normale pour cela la connaissance de la distribution de ces dernière présente un étape et condition nécessaire pour appliquer d'autre analyse. La conformité de données avec les lois de probabilité s'appelle les tests d'ajustement et pour les conformé avec la loi normale nous obtenons les tests de normalité. Ce travail, porte surles tests de normalité, ou nous somme présenté quelques méthodes graphiques et théoiques. Les méthode graphiques comme l'histogramme de fréquence, la boite à moustache, le graphe quantile- quantile et la droite de Henry . La méthode théorique comme les tests de Kolmogorov-Smirnov, test de Shapiro -Wilk,test de Lilliefors, testd'Anderson –Darling, test de Cramer-Von Mises et test de Jarque-Bera.

Mots clés : Test de normalité, test d'hypothèse, hypothèse nulle, hypothèse alternative, la statistique, région critique, P-valeur.

ملخص:

يعتبر التوزيع الطبيعي ذا اهمية بالغة في الاحصاء والاحتمالات وذلك نظرا لان أغلبية المعطيات تتبع التوزيع الطبيعي ولهذا فإن معرفة توزيع هذه المعطيات يعتبر مرحلة وشرط اساسي لتطبيق طرق اخرى. توافق المعطيات مع القوانين الاحتمالية يسمى اختبار المطابقة وبالنسبة للتوزيع الطبيعي فهو اختبار الحالة الطبيعية للمعطيات.هذا العمل, يحمل طرق اختبار الحالة الطبيعية, كيفية التحقق من طبيعية المعطيات باستخدام طريقة الرسم كتردد مدرج التكراري, صندوق الشارب, الرسم البياني الكمي و خط هنري . الطريقة النظرية كاختبار كولموغوروف-سميرنوف, اختبار شابيرو-ويلك, اختبار ليليفورس, اختبار اندرسون-دارلينغ, اختبار كرامر-فون ميس و اختبار جارك- بيرا .

الكلماتالمفتاحية :

اختبار الحالة الطبيعية, اختبار الفرضية, فرضية المعدومة, فرضية بديلة, احصائيات, منطقة الحرجة, قيمة -P

Abstract

The normal law plays a very important role in statistics also in probability because most of the data in reality follow the normal distribution for that the knowledge of the distribution of the latter presents a necessary step and condition to apply other analysis. The conformity of data with the probability laws is called the goodness-of-fitGraphical methods like frequency histogram, boxplot, quantile-quantile graph, and Henry's line. Theoretical method tests and for conforming them to the normal distribution we obtain the tests of normality. This work focuses on normality tests, where we have presented some graphic and theoretical methods. such asKolmogorov-Smirnov tests, Shapiro -Wilk test,Lillie fors test, Anderson -Darling test,Cramer-Von Mises test and Jarque-Bera test.

Key words:

Normality test-hypothesis test-null hypothesis-alternative hypothesis-significance level-critical region-P-value.