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

Achour Chams

Titre :

Tests de normalité et applications

Membres du Comité d'Examen :

Dr.	Sayah Abdallah	UMKB	Président
Dr.	Roubi Afaf	UMKB	Encadreur
Dr.	Abdelli Jihane	UMKB	Examinatrice

Juin 2021

Dédicace

Je dédie cet humble travail à :

À mes très chers parents Brahim et Fatima pour leur soutien et leurs précieux conseils, pour toute leur assistance et leur présence dans ma vie.

À mes très chers frères et sœurs.

À tous mes proches.

À tous mes amis, particulièrement Haykal Ahlem et Ghodbane Sara Rayane.

À mes amis et mes collègues de la promotion 2021

«Mathématiques».

REMERCIEMENTS

J'exprime d'abord mes profonds remerciements à "**ALLAH**"

qui m'a donné la volonté et le

courage pour la réalisation de ce travail.

J'exprime mes profondes gratitudes à mes parents.

Je tiens à exprimer ma profonde gratitude à mon encadreur Dr

Roubi Afaf et je le remercie aussi de ses remarques importantes

et de son suivi permanent de mon travail.

Je veux exprimer aussi tout mon respect aux membres du jury

Dr Sayah Abdallah et Dr Abdelli Jihane

qui ont acceptés d'évaluer et de juger mon travail.

Mes remerciements vont aussi à tous les enseignants du département de Mathématiques

qui ont contribué à ma formation.

Table des matières

Table des figures

Liste des tableaux

Introduction

Les tests de normalité sont des cas particuliers des tests statistiques prennent une place importante en statistique. En effet, de nombreux tests supposent la normalité des distributions pour être applicables. En toute rigueur, il est indispensable de vérifier la normalité avant d'utiliser les tests. Cependant de nombreux tests sont suffisamment robustes pour être utilisables même si les distributions s'écartent de la loi normale.

L'objectif de ce mémoire est l'étude des techniques statistiques destinées à examiner la compatibilité d'une distribution empirique avec la loi normale. On parle également de test d'adéquation à la loi normale.

Ce mémoire est composé de deux chapitres

– Premier chapitre : On a étudié la loi normale et ses propriétés fondamentales puis on a présenté les notions principales d'un test d'hypothèse.

– Deuxième chapitre : Ce chapitre est consacré aux méthodes statistiques qui a pour objectif de vérifier l'hypothèse de normalité des données, dont on a présenté des méthodes graphiques telles que l'histogramme, droite de Henry,..., et des méthodes théoriques comme le test de Kolmogorov-Smirnov, le test de Shapiro-Wilk,..., etc.

Chapitre 1

Loi normale et test d'hypothèses

Dans ce premier chapitre, on va étudier dans la première section une loi de probabilité très importante en statistique qui est la loi normale dont ses différentes propriétés seront données puis on va donner les notions importantes dans la théorie des tests d'hypothèses.

1.1 La loi normale

1.1.1 Quelques définitions et propriétés

Définition 1.1.1 *Soit X une variable aléatoire réelle suit une loi normale (ou loi gaussienne) d'espérance μ et d'écart-type σ (et on note $X \backsim \mathcal{N}(\mu, \sigma^2)$), alors la variable aléatoire réelle X admet pour densité de probabilité la fonction $f(x)$ définie par*

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

Proposition 1.1.1 *Si* $X \backsim \mathcal{N}(\mu, \sigma^2)$ *alors*

1. La fonction de répartition de X *est*

$$\begin{aligned} F_X(t) &= P(X \leq t). \\ &= \int_{-\infty}^{t} f(x)dx. \\ &= \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-1}{2}(\frac{x-\mu}{\sigma})^2)dx. \end{aligned}$$

2. La moyenne de X *est* $E(X) = \mu$ *et la variance de* X *est* $Var(X) = \sigma^2$.

Des exemples de la densité de la loi normale $\mathcal{N}(\mu, \sigma^2)$ selon les valeurs de la moyenne et de la variance sont donnés aux figures 1.1 et 1.2.

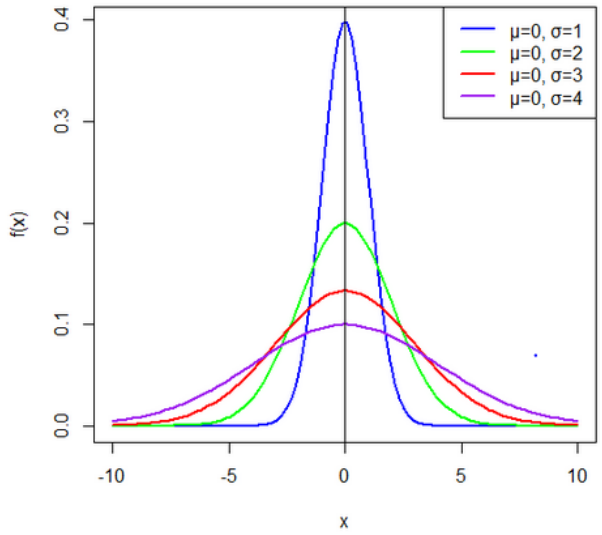

FIG. 1.1 – Densités de la loi normale de moyenne $\mu = 0$ et d'écart-types $\sigma = 1, 2, 3, 4$.

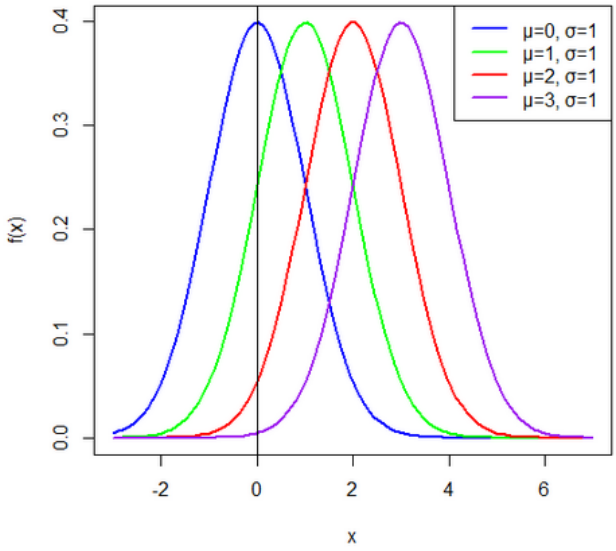

FIG. 1.2 – Densités de la loi normale d'écart-type $\sigma = 1$ et de moyennes $\mu = 0, 1, 2, 3$.

Loi normale centrée réduite

Soit Z une v.a de loi normale. On dit que Z suit une loi normale centrée réduite si sa moyenne $\mu = 0$ et son écart-type $\sigma = 1$. La loi sera donc notée $\mathcal{N}(0, 1)$.

La fonction de densité sera donnée par

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-z^2}{2}), \qquad z \in \mathbb{R},$$

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

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

Propriété 1.1.1 $\Phi(z) = 1 - \Phi(-z), \qquad z \in \mathbb{R}.$

Les valeurs $\Phi(z)$ sont données dans la table statistique de loi normale centrée réduite $\mathcal{N}(0, 1)$.

Exemple 1.1.1 *1.* $\Phi(0) = P(Z \leq 0) = 0.5$ *;*

2. $\Phi(1.53) = P(Z \leq 1.53) = 0.937$ *;*

3. $\Phi(-1.53) = P(Z \leq -1.53) = 1 - 0.937 = 0.063$.

Théorème 1.1.1 *Soit* X *une v.a suit une loi normale* $\mathcal{N}(\mu, \sigma^2)$ *alors la v.a*

$$Z = \frac{X - \mu}{\sigma} \backsim \mathcal{N}(0, 1),$$

donc

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

Exemple 1.1.2 *Soit* X *une v.a suit la loi normale* $\mathcal{N}(3, 16)$*, on va calculer*

1. $P(X \leq 3.8)$.

2. $P(6 \leq X \leq 7)$.

On utilise la table de la loi normale $\mathcal{N}(0, 1)$

1. $P(X \leq 3.8) = \Phi(\frac{3.8-3}{4}) = \Phi(0.2) = 0.5793$.

2. $P(6 \leq X \leq 7) = \Phi(\frac{7-3}{4}) - \Phi(\frac{6-3}{4}) = 0.8413 - 0.7734 = 0.0679$.

1.1.2 Théorème centrale limite

Théorème 1.1.2 *Soit une suite* (X_n) *de variables aléatoires définies sur le même espace de probabilité, suivant la même loi* D *et dont l'espérance et l'écart-type communes existent et soient finis* $(\sigma \neq 0)$*. On suppose que les* (X_n) *sont indépendantes. Considérons la somme* $S_n = X_1 + \cdots + X_n$*. Alors l'espérance de* S_n *est* $n\mu$ *et son écart-type vaut* $\sigma\sqrt{n}$ *et* $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ *converge en loi vers une variable aléatoire normale centrée réduite on note* $(Y_n \backsim \mathcal{N}(0, 1))$ *[8].*

1.2 Test d'hypothèses

Définition 1.2.1 *Un test d'hypothèses est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou accepter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un échantillon de données.*

Un test est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon.

Remarque 1.2.1 *Les tests d'hypothèses ont pour buts de*

– clarifier et définir le cadre rigoureux de ces études.

– fournir un formalisme précis pour toutes les situations.

– savoir si les différences mises en jeu sont importantes ("significatives" pour un seuil donné) ou non [5].

1.2.1 Hypothèse nulle et hypothèse alternative

Définition 1.2.2 ***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 [4].*

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 [4].*

Remarque 1.2.2 *1. Une hypothèse doit spécifier une valeur, disons θ_0 pour un paramètre θ de la population. On testera donc : $H_0 : \theta = \theta_0$.*

Une possibilité classique pour l'hypothèse alternative est : $H_1 : \theta \neq \theta_0$, qui teste chaque côté de l'égalité (on parlera de test bilatéral).

Mais on peut écrire également un autre choix d'hypothèse :

$H_0 : \theta \geq \theta_0$ *parfois noté encore* $H_0 : \theta = \theta_0$.

et l'hypothèse alternative correspondante sera : $H_1 : \theta < \theta_0$, *qui teste un seul côté de l'égalité (on parlera de test unilatéral).*

2. Le raisonnement inverse peut être formulé l'hypothèse suivant

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

3. $(H_0 : \theta \leq \theta_0$ *ou* $H_0 : \theta \geq \theta_0)$ *parfois noté encore* $(H_0 : \theta = \theta_0)$ *[5].*

1.2.2 Test bilatéral et test unilatéral

Le choix entre le test bilatéral et unilatéral dépend de la nature des données, du type d'hypothèse que l'on désire à contrôler, des affirmations que l'on peut admettre concernant la nature des populations étudiées (normalité, égalité des variances) [10].

Test d'hypothèses unilatéral

Nous considérons deux situations du type unilatéral

$$\begin{cases} H_0 : & \theta \leq \theta_0 \\ H_1 : & \theta > \theta_0 \end{cases}, \quad \text{ou} \quad \begin{cases} H_0 : & \theta \geq \theta_0 \\ H_1 : & \theta < \theta_0 \end{cases},$$

où θ est un paramètre de dimension 1 ($\Theta \subseteq \mathbb{R}$). Hypothèse nulle et hypothèse alternative sont multiples. De telles situations se rencontrent lorsque l'on s'intéresse uniquement à juger si le paramètre θ dépasse un certain seuil. L'hypothèse nulle est dite unilatéral et par extension on parle aussi de test unilatéral du fait que la région de rejet est usuellement de la forme $T > c$ ou $T < c$, T étant la statistique de test [9].

Tests d'hypothèses bilatérales

Nous considérons deux situations du type bilatéral

$$\begin{cases} H_0 : \ \theta = \theta_0 \\ H_1 : \ \theta \neq \theta_0 \end{cases}, \quad \text{ou} \quad \begin{cases} H_0 : \ \theta_1 \leq \theta \leq \theta_2 \\ H_1 : \ \theta < \theta_1 \text{ ou } \theta > \theta_2 \end{cases},$$

où θ est un paramètre de dimension 1 ($\Theta \subseteq \mathbb{R}$). La première situation est fréquente lorsque θ représente en fait un écart entre paramètres de deux populations. La seconde teste si le paramètre est situé dans un intervalle de tolérance acceptable. L'appellation de bilatéral se réfère au fait que l'alternative est située de part et d'autre de l'hypothèse nulle [9].

1.2.3 Variable de décision et niveau de signification

Variable de décision : 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 [5].

Niveau de signification : Le niveau de signification (ou seuil de signification) est égale un risque de première espèce maximum. On le note par α.

Les valeurs usuelles de niveau de signification sont $1\%, 5\%$ *et* 10%.

1.2.4 Région critique et risque d'erreur

La région de rejet et région critique

Définition 1.2.3 *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 [5]. Elle est notée par $\overline{W}$ *et est définie par*

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

Risque d'erreur

Définition 1.2.4 *(**Erreur**) 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.*

Définition 1.2.5 *(***Risque***) 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.*

Exemple 1.2.1 *Si l'on cherche à tester l'hypothèse qu'une pièce de monnaie n'est*

pas « truquée », nous allons adopter la règle de décision suivante :

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

– *on accepte* H_0 *si* $X \in [40; 60]$ *;*

– *on rejette* H_0 *si* $X \notin [40; 60]$ *; donc soit* $X < 40$ *ou* $X > 60$*, avec* X *« nombre de faces » obtenus en lançant* 100 *fois la pièce.*

On suppose que la probabilité d'obtenir face est de p *c'est une pièce truquée, et* $(1-p)$ *d'obtenir pile c'est une pièce n'est pas truquée et on a* $p = (1-p) = \frac{1}{2}$ *donc les hypothèses à tester sont* $\begin{cases} H_0 : & p = \frac{1}{2} \\ H_1 : & p \neq \frac{1}{2} \end{cases}$.

La v.a X *: "nombre de faces obtenues" suit une loi Binomiale tel que*

$X \backsim \mathcal{B}(100, \frac{1}{2})$*, où* $n = 100$ *: "Nombre de lancés".*

p *: Probabilité du succès" obtenir face", p* $= \frac{1}{2}$ *la pièce n'est pas truquée.*

Alors

1. Quel est le risque d'erreur de première espèce ?

2. Quel est le risque d'erreur de deuxième espèce, tel que : $p = 0.6$ *?*

Solution 1.2.1 *1. Le risque d'erreur de première espèce* α *c'est*

On a : $X \backsim \mathcal{B}(100, \frac{1}{2})$*, 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

$$\begin{aligned}
1-\alpha &= P(X \in [40,60]) \\
&= P(40 \leq X \leq 60) \\
&= P(\frac{40-50}{5} \leq Z \leq \frac{60-50}{5}) \\
&= P(-2 \leq Z \leq 2) = P(Z \leq 2) - P(Z \leq -2) \\
&= P(Z \leq 2) - (1 - P(Z \leq 2) = 2P(Z \leq 2) - 1 \\
&= 2 \times 0.9772 - 1 = 0.9544
\end{aligned}$$

$$1-\alpha = 0.9544 \Longleftrightarrow \alpha = 1 - 0.9544 = 0.0456 \backsimeq 0.05.$$

Alors le risque d'erreur de première espèce est $\alpha = 0.05$.

2. Le risque d'erreur de deuxième espèce c'est β

On a : $X \backsim \mathcal{B}(100, 0.6)$, *alors* $E(X) = np = 100 \times 0.6 = 60$,

et $Var(X) = np(1-p) = 100 \times 0.6 \times 0.4 = 24$.

Le risque d'erreur de deuxième espèce β *c'est*

$$\beta = P(accepter\ H_0 / H_0\ est\ fausse) = P(40 \leq X \leq 60).$$

On a Z *suit la loi* $\mathcal{N}(0,1)$ *tq :* $Z = \frac{X - E(X)}{\sqrt{Var(X)}}$.

Alors

$$\begin{aligned}
\beta &= P(\frac{40-60}{\sqrt{24}} \leq Z \leq \frac{60-60}{\sqrt{24}}) \\
&= P(-4.08 \leq Z \leq 0) = P(Z \leq 0) - P(Z \leq -4.08) \\
&= 0.5 - (1 - 0.9999) \backsimeq 0.50.
\end{aligned}$$

Donc : le risque d'erreur de deuxième espèce est $\beta = 0.5$.

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".

1.2.5 Règle de décision

Soit H_0 et H_1 deux hypothèses dont une et une seule est vraie. La décision aboutira à choisir H_0 ou H_1. Il y a donc quatre cas possibles qu'ils sont résumées dans le tableau suivant

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

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

Définition 1.2.6 *La probabilité de rejeter H_0 alors qu'elle est vraie vaut α et est appelé niveau du test (ou seuil). La probabilité de rejeter une fausse hypothèse nulle est $(1-\beta)$ qui est appelée la puissance du test.*

1.2.6 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 significativité α et p-valeur (P) en vue de savoir accepter ou rejeter H_0

– Si P $\leq \alpha$ on va rejeter l'hypothèse H_0.

– Si P $> \alpha$ on va accepter 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 Tests paramétriques

Définition 1.3.1 *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. Soit* $(X_1, X_2, ..., X_n)$ *un échantillon issu d'une v.a de la loi* $P(\theta \in \Theta)$. *On a considéré* Θ_0 *et* Θ_1 *avec :* $\Theta_0 \cup \Theta_1 = \Theta$ *et* $\Theta_0 \cap \Theta_1 = \phi$.

Donc on a deux hypothèses à 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 dits "tests à un échantillon". Ils ont pour but de vérifier si un échantillon peut être considéré comme représentatif de la population dont il est extrait. On étudie une variable quantitative* X *et on cherche à établir si les observations sont en accord avec la loi théorique de cette variable [11].*

En général, il s'agit de tester si un paramètre (tel que la moyenne, la fréquence ou la variance) calculé dans l'échantillon est conforme à sa valeur au niveau de la population. Ceci suppose que la loi théorique du paramètre est connue au niveau de la population [11].

Soit $X_1, X_2, ..., X_n$ *un échantillon de taille* n *d'une v.a* $X \backsim \mathcal{N}(\mu, \sigma^2)$ *de moyenne* μ *et de variance* σ^2*et de proportion* p, *pour un niveau critique* α *fixée, les tests de conformité sont résumés dans le tableau suivant [3]*

L'hypothèse	Cas possible	La statistique	La région critique
$\begin{cases} H_0 : \ \mu = \mu_0 \\ H_1 : \ \mu \neq \mu_0 \end{cases}$.	σ^2 connue	$U = \frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}$	$\lvert U\rvert > u_{\frac{\alpha}{2}}$
	σ^2 inconnue $n < 30$	$T = \frac{\sqrt{n}(\overline{X}-\mu_0)}{S}$	$\lvert T\rvert > t_{\frac{\alpha}{2}}(n-1)$
	σ^2 inconnue $n \geq 30$	$U = \frac{\sqrt{n}(\overline{X}-\mu_0)}{S}$	$\lvert U\rvert > u_{\frac{\alpha}{2}}$
$\begin{cases} H_0 : \ \sigma^2 = \sigma_0^2 \\ H_1 : \ \sigma^2 \neq \sigma_0^2 \end{cases}$.	μ inconnue	$\mathcal{X}_0^2 = (n-1)\frac{S^2}{\sigma_0^2}$	$\mathcal{X}_0^2 \notin [\mathcal{X}^2_{1-\frac{\alpha}{2}}(n-1);$ $\mathcal{X}^2_{\frac{\alpha}{2}}(n-1)]$
	μ connue	$\mathcal{X}_1^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$	$\mathcal{X}_1^2 \notin [\mathcal{X}^2_{1-\frac{\alpha}{2}}(n);\ \mathcal{X}^2_{\frac{\alpha}{2}}(n)]$
$\begin{cases} H_0 : \ p = p_0 \\ H_1 : \ p \neq p_0 \end{cases}$.	$X \backsim \mathcal{B}(n,p)$ et $n \geq 30$	$U = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\lvert U\rvert > u_{\frac{\alpha}{2}}$

TAB. 1.2 – Résumé sur le test de conformité.

Remarque 1.3.1 *1. La statistique U suit la loi normale $\mathcal{N}(0,1)$.*

2. $u_{\frac{\alpha}{2}}$: Le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi normale centrée réduite.

3. La Statistique T suit la loi de Student de $(n-1)$ ddl.

4. $t_{\frac{\alpha}{2}}$: Le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi Student à $(n-1)$ ddl.

5. La Statistique $\mathcal{X}_0^2$ suit la loi de Khi-deux de $(n-1)$ ddl.

6. La Statistique $\mathcal{X}_1^2$ suit la loi de Khi-deux de n ddl.

7. L'écart type corrigé S égale à $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}$.

1.3.2 Test d'homogénéité

Définition 1.3.3 *Il s'agit de comparer deux populations à l'aide d'un paramètre donné tel que la moyenne, la variance et la fréquence,... etc.*

Soient X_1, X_2 deux v.a de moyennes μ_1, μ_2 et de variances σ_1^2, σ_2^2.

Soient $X_{1,1}, X_{1,2}, ..., X_{1,n}$ et $X_{2,1}, X_{2,2}, ..., X_{2,n}$ 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. Soit

$X_i \backsim \mathcal{N}(\mu_i, \sigma_i)$, $i = 1, 2$. Les tests des homogénéités sont résumés dans le tableau suivant [3]

L'hypothèse	Cas possible	La statistique	La région critique
$\begin{cases} H_0 : \ \mu_1 = \mu_2 \\ H_1 : \ \mu_1 \neq \mu_2 \end{cases}$.	σ_i^2 connues. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (où σ^2 inconnue)	$U = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $T = \frac{\overline{X_1} - \overline{X_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$\lvert U \rvert > u_{\frac{\alpha}{2}}$ $\lvert T \rvert > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$
$\begin{cases} H_0 : \ \sigma_1^2 = \sigma_2^2 \\ H_1 : \ \sigma_1^2 \neq \sigma_2^2 \end{cases}$.	σ_1^2 et σ_2^2 sont inconnues	$F = \frac{S_1^2}{S_2^2}$	$F > \mathcal{F}_{1-\frac{\alpha}{2}}(n_1 - 1;$ $n_2 - 1)$
$\begin{cases} H_0 : \ p_1 = p_2 = p \\ H_1 : \ \ p_1 \neq p_2 \end{cases}$.	$X_i \backsim \mathcal{B}(p_i), i = 1, 2$ n_1, n_2 très grands	$U = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$\lvert U \rvert > u_{\frac{\alpha}{2}}$

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

Remarque 1.3.2 *1. La statistique U suit la loi normale $\mathcal{N}(0, 1)$.*

2. $u_{\frac{\alpha}{2}}$: Le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi normale centrée réduite.

3. La statistique T suit la loi de Student à $(n_1 + n_2 - 2)$ ddl.

4. $t_{\frac{\alpha}{2}}$: Le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi de Student à $(n_1 + n_2 - 2)$ ddl.

5. La statistique F suit la loi de Fisher à $(n_1 - 1; n_2 - 1)$ ddl.

6. $\mathcal{F}_{1-\frac{\alpha}{2}}$: Le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi de Fisher à $(n_1 - 1; n_2 - 1)$ ddl.

7. La Statistique $\mathcal{X}_0^2$ suit la loi de Khi-deux de $(n - 1)$ ddl.

8. La variance corrigée commune S^2 égale à $S^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ où s_1^2, s_2^2 sont les estimateurs de σ_1^2 et σ_2^2 respectivement.

1.4 Tests non paramétriques

Définition 1.4.1 *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 [4].*

• *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 [4]. Parmi les tests d'adéquation on peut citer le test d'adéquation de Khi-deux.*

Test d'adéquation de Khi-deux

Soit X une variable aléatoire de loi $\mathcal{L}$ (le plus souvent inconnue). On souhaite tester l'ajustement de cette loi à une loi connue $\mathcal{L}_0$ ($Poisson$, $Exponentielle$, $normale$,... etc) retenue comme étant un modèle convenable [8].

On teste donc l'hypothèse

$$\begin{cases} H_0 : & \mathcal{L} = \mathcal{L}_0 \\ H_1 : & \mathcal{L} \neq \mathcal{L}_0 \end{cases}.$$

Les n observations de X sont partagées en k classes. On désigne par O_i l'effectif observé de la classe i. Ainsi $\sum_i O_i = n$ [8].

Pour chaque classe, l'effectif théorique est défini

$$C_i = n.P(X \in Classe_i | X \longrightarrow \mathcal{L}_0).$$

On calcule la valeur $\mathcal{X}_c^2 = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$. On compare cette valeur à la valeur théorique $\mathcal{X}_\alpha^2$ lue dans la table du $\mathcal{X}^2$ à $v = k - 1 - r$ degrés de liberté où r est le nombre de paramètres de la loi $\mathcal{L}_0$ qu'il a fallu estimer. On rejette H_0 lorsque $\mathcal{X}_c^2 > \mathcal{X}_\alpha^2$ [8].

Chapitre 2

Tests de normalité

En statistique, les tests servent à vérifier si des données réelles suivent une loi normale ou non sont appelés tests de normalité. Les tests de normalité sont des cas particuliers des tests d'adéquation ou d'ajustement, appliqués à une loi normale. Ces tests prennent une place importante en statistique. Dans ce chapitre on va donner les techniques statistiques visant à tester la normalité des données qu'on peut les diviser en : des méthodes graphiques (Histogramme de fréquence, *Q-Q plot*,... etc), et des méthodes théoriques (test de Kolmogorov-Smirnov, test de Shapiro-Wilk, test de Lilliefors ,... etc).

Définition 2.0.2 *Le test de normalité est un test non paramétrique des hypothèses :*

$$\begin{cases} H_0 : & \textit{La loi de la v.a } X \textit{ est une normale} \\ H_1 : & \textit{La loi de la v.a } X \textit{ n'est pas une normale} \end{cases}.$$

2.1 Méthodes graphiques

2.1.1 Histogramme de fréquence

L'outil graphique le plus simple est l'histogramme de fréquence. Il 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.

Le nombre k d'intervalles est défini de manière arbitraire, dans d'autres il est paramétrable. Une règle simple pour définir le bon nombre d'intervalles est d'utiliser la règle $k = \log(n)$, n est la taille de l'échantillon. Il possible de visualiser la forme de la distribution des données à analyser en les représentant sous forme d'histogramme avec une courbe représentant une loi normale [1].

Programmation en code R

```
x=rnorm(100)
y=rnorm(150)
par(mfrow=c(1,2))
hist(x,main="Histogramme de x",ylab="fréquence", prob=T)
curve(dnorm(x, mean(x), sd(x)),add=TRUE)
hist(y,main="Histogramme de y",ylab="fréquence", prob=T)
curve(dnorm(x, mean(y), sd(y)),add=TRUE)
```

Résultat de la commande

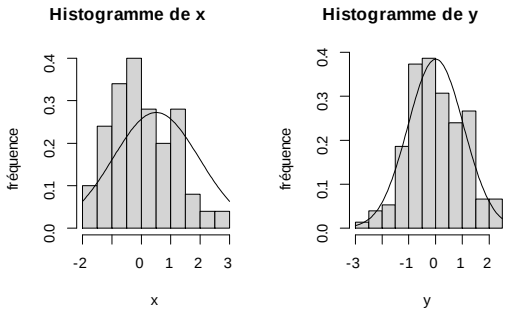

FIG. 2.1 – Histogramme de fréquence avec densité de la loi normale de x et y.

2.1.2 Boîte à moustache

La boîte à moustaches, en anglais box-plot, est un outil graphique très pratique représentant une distribution empirique à l'aide de quelques paramètres de localisation : la médiane (M), le 1er (Q_1) et 3ème (Q_3) quartile [1].

La boîte correspond à la partie centrale de la distribution : la moitié des valeurs comprises entre le premier et le troisième quartile Q_1 et Q_3. Les moustaches s'étendent de part et d'autre de la boîte jusqu'aux valeur suivantes : à gauche jusqu'à $Q_1 - 1.5(Q_3 - Q_1)$ s'il existe des valeurs encore plus petites, sinon jusqu'à la valeur minimale à droite jusqu'à $Q_1 + 1.5(Q_3 - Q_1)$ s'il existe des valeurs au-delà, sinon jusqu'à la valeur maximale. Les valeurs au-delà des moustaches repérées par des noms des valeurs hors norme éventuellement suspectes ou aberrantes mais pas nécessairement [2].

Programmation en code R

x=rnorm(100)

y=rnorm(150)

summary(x)

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

-2.0129 -0.4573 0.1266 0.1179 0.6582 2.2643

summary(y)

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

-3.06949 -0.82961 -0.04854 -0.05132 0.69633 2.44189

boxplot(x,y,names=c("x","y"))

Résultat de la commande

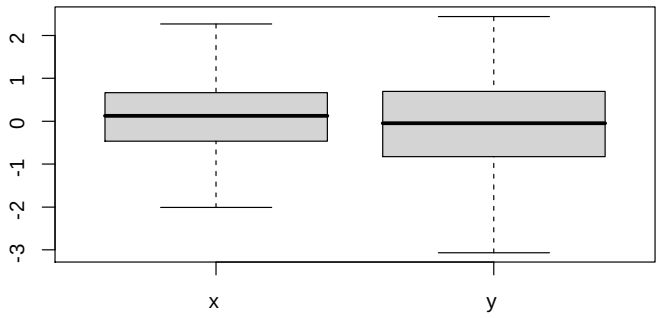

FIG. 2.2 – Boîte à moustache de x et y.

2.1.3 Q-Q plot et 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 de données.

Les échantillons ne sont pas forcément de même taille. Il se peut également, et c'est ce qui nous intéresse dans le cas présent, qu'un des ensembles de données soient

générées à partir d'une loi de probabilité qui sert de référentiel [1].

Concrètement, 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 $F_i = \frac{i-0.375}{n+0.25}$;

3. nous calculons les quantiles successifs $z_{*(i)}$ d'ordre F_i en utilisant l'inverse de la loi normale centrée et réduite ;

4. enfin, les données initiales n'étant pas centrées et réduites, nous dé-normalisons les données en appliquant la transformation $x_{*(i)} = z_{*(i)} \times S + \overline{X}$ [1], où $\overline{X}$ est l'estimateur de la moyenne (moyenne empirique) et S est l'estimateur de l'écart type σ et sont

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

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 [8].

Principe : On représente les quantiles théoriques en fonction des quantiles observés (Diagramme Q-Q). Si X est une variable gaussienne de moyenne $\overline{x}$ et de variance σ^2 et si Z est une variable aléatoire de loi normale centrée réduite, on a les égalités suivantes

$$P(X < x_i) = P(\frac{X - \overline{x}}{\sigma} < \frac{x_i - \overline{x}}{\sigma}) = P(Z < y_i) = \Phi(y_i),$$

où $y_i = \frac{x_i - \overline{x}}{\sigma}$.

(On note Φ la fonction de répartition de la loi normale centrée réduite) [8].

Pour chaque valeur x_i de la variable X, on peut calculer $P(X < x_i)$ puis en déduire,

à l'aide d'une table de la fonction Φ, y_i tel que : $\Phi(y_i) = P(X < x_i)$.

Si la variable est gaussienne, les points de coordonnées $(x_i ; y_i)$ sont alignés sur la droite d'équation $y = \frac{x-\overline{x}}{\sigma}$ [8].

Programmation en code R

```
x=rnorm(100)
y=rnorm(150)
par(mfrow=c(1,2))
qqnorm(x, datax=TRUE)
qqline(x, datax=TRUE)
qqnorm(y, datax=TRUE)
qqline(y, datax=TRUE)
```

Résultat de la commande

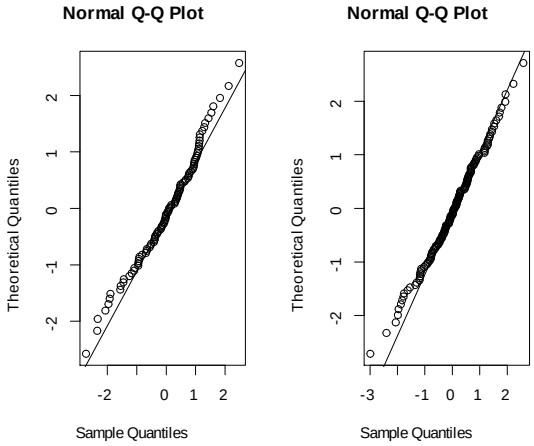

FIG. 2.3 – Q-Q plot et droite de Henry pour x et y.

2.2 Méthodes théoriques

2.2.1 Test de Kolmogorov-Smirnov

L'objectif est d'établir la plausibilité de l'hypothèse selon laquelle l'échantillon a été prélevé dans une population ayant une distribution donnée. Le test de Kolmogorov est "non paramétrique" : il ne place aucune contrainte sur la distribution de référence, et ne demande pas qu'elle soit connue sous forme analytique (bien que ce soit pourtant le cas le plus courant). Etant donnés

1. Un échantillon de taille n d'observations d'une variable.

2. Et une fonction de répartition de référence F(x) [8].

Soit $X_1, ..., X_n$ un n-échantillon d'une v.a X de loi P absolument continue par rapport à la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ inconnue. On note F_n la distribution empirique associée à X.

Théorème 2.2.1 *$(Glivento - Cantelli)$*

Soit $(X_i)_{i\geq 1}$ une suite des variables aléatoires i.i.d de fonction de répartition F. On pose

$$\forall t \in \mathbb{R},\ F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq t\}}.$$

Alors on a

$$\sup_{t\in\mathbb{R}} |F_n - F(t)| \longrightarrow 0\ Ps\ [4].$$

Définition 2.2.1 *La statistique de KS est défini par la distance en norme infinie de la fonction de répartition empirique F_n, et la fonction de répartition F*

$$KS = D_{KS}(P, P_0) = \sup_{t\in\mathbb{R}} |F_n - F(t)|\ .$$

Propriété 2.2.1 *Si $(x_{(1)}, x_{(2)}, ..., x_{(n)})$ est la statistique d'ordre associée à l'échan-*

tillon X alors

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

On rejet H_0 si $\sqrt{n}D_{KS} > d_{n,\alpha}$ où $d_{n,\alpha}$ est le quantile théorique lu à partir la table de Kolmogorov-Smirnov [4].

Programmation en code R

On compare les p-value de tout les tests par $\alpha = 0.05$, c-à-d

$$\begin{cases} \text{si p-value} > 0.05 \text{ on accepte } H_0 \\ \text{si p-value} < 0.05 \text{ on rejette } H_0 \end{cases}.$$

x=rnorm(100)

y=rnorm(100)

ks.test(x,y)

Résultat de la commande

Two-sample Kolmogorov-Smirnov test

data : x and y

D = 0.06, p-value = 0.9938

alternative hypothesis : two-sided

Commentaire : On remarque que la p-value est supérieure au niveau α alors on accepte l'hypothèse H_0. Alors on accepte l'hypothèse de normalité de l'échantillon au risque $\alpha = 5\%$.

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 ($n \leq 50$). La statistique du test s'écrit

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

où

• $x_{(i)}$ correspond à la série des données triées ;

• $\left[\frac{n}{2}\right]$ est la partie entière du rapport $\frac{n}{2}$;

• a_i sont des constantes générées à partir de la moyenne et de la matrice de variance covariance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont fournies dans des tables spécifiques [6].

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 quantiles empiriques obtenues à partir des données [6].

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

$$R.C : W < Wcrit.$$

Les valeurs seuils W_{crit} pour différents risques 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 [6].

Programmation en code R

x=rexp(10,2)

shapiro.test(x)

Résultat de la commande

Shapiro-Wilk normality test

data : x

W = 0.80005, p-value = 0.01452

Commentaire : On remarque que la p-value est inférieure au niveau α alors on rejette l'hypothèse H_0. Alors on rejette l'hypothèse de normalité de l'échantillon au risque $\alpha = 5\%$.

2.2.3 Test de Lilliefors

Ce test est une variante du test de Kolmogorov-Smirnov, sous l'hypothèse de normalité (à chercher à tester $H_0 : P \backsim Gaussienne$), où les paramètres μ et σ de la loi sont estimés à partir des données.

La statistique du test est

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

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

La table des valeurs critiques D_{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	D_{crit}
0.10	$\frac{0.885}{\sqrt{n}}$
0.05	$\frac{0.886}{\sqrt{n}}$
0.01	$\frac{1.031}{\sqrt{n}}$

TAB. 2.1 – Les valeurs critiques de test Lilliefors.

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

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

Programmation en code R

x=rnorm(100)

lillie.test(x)

Résultat de la commande

Lilliefors (Kolmogorov-Smirnov) normality test

data : x

D = 0.059021, p-value = 0.5318

Commentaire : On constate qu'au seuil de risque $\alpha = 5\%$, on accepte l'hypothèse de normalité de l'échantillon car p-value$> \alpha$.

2.2.4 Test de Anderson-Darling

Le test de 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 œuvre d'un test paramétrique (comparaison de moyenne, de variances, etc.) que le test de Lilliefors [1].

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

Concernant l'adéquation à la loi normale, la statistique du test s'écrit [1]

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

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

Une correction est recommandée pour les petits effectifs, cette statistique corrigée est également utilisée pour calculer la *p-value* :

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

Les valeurs critiques A_{crit} pour différents niveaux de risques sont résumées dans le tableau suivant, elles ont été produites par simulation et ne dépendent pas de l'effectif de l'échantillon [1]

α	A_{crit}
0.10	0.631
0.05	0.752
0.01	1.035

TAB. 2.2 – Les valeurs critiques de test d'Andarson Darling.

L'hypothèse de normalité est rejetée lorsque la statistique A prend des valeurs trop élevées

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

Programmation en code R

x=rnorm(100)

ad.test(x)

Résultat de la commande

Anderson-Darling normality test

data : x

A = 0.2331, p-value = 0.7925

Commentaire : Puisque p-value> α ($0.7925 > 0.05$) on ne peut pas rejeter l'hypothèse de normalité de l'échantillon au risque $\alpha = 5\%$.

2.2.5 Test de Cramer-Von Mises

Le test de Cramer-Von Mises ou critère de Cramer-Von Mises est un test statistique utilisé pour évaluer la qualité de l'ajustement d'une fonction de répartition notée F comparé à une fonction de répartition empirique notée F_{emp}.

Ce test est nommé en l'honneur d'Harald Cramer etRichard Von Mises. Ce test est également une alternative au test de Kolmogorov-Smirnov.

La statistique du test s'écrit

$$W_n^2 = n \int_{-\infty}^{+\infty} |F_{emp}(x) - F(x)|^2 \, dF(x).$$

$$W_n^2 = \sum_{i=1}^{n} (F(_{(x_i)}) - \frac{2i-1}{2n})^2 + \frac{1}{12n}.$$

Région critique du test : On rejette H_0 si

$$R.C : W_n^2 \geq W_{crit}.$$

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

Programmation en code R

x=rexp(10,2)

cvm.test(x)

Résultat de la commande

Cramer-von Mises normality test

data : x

W = 0.28119, p-value = 0.0003781

Commentaire : On remarque que la p-value est inférieure au niveau α, alors on rejette l'hypothèse H_0. C'est à dire, on rejette l'hypothèse de normalité de l'échantillon au risque $\alpha = 5\%$.

2.2.6 Test de Jarque-Bera

Le test de normalité de Jarque-Bera est 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. La formulation est très simple. Il ne devient réellement intéressant que lorsque les effectifs sont élevés [1].

Les coefficients d'asymétrie et d'aplatissement :

1) Les coefficients d'asymétrie : En théorie des probabilités et statistique, le coefficient d'asymétrie (skewness en anglais) correspond à une mesure de l'asymétrie

de la distribution d'une variable aléatoire réelle.

Étant donnée une variable aléatoire réelle X de moyenne μ et d'écart type σ, on définit son coefficient d'asymétrie comme le moment d'ordre trois de la variable centrée réduite

$$\beta_1 = \mathbb{E}[(\frac{X-\mu}{\sigma})^3],$$

lorsque cette espérance existe. On a donc : $\beta_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$, avec μ_i les moments centrés d'ordre i.

Forme de la distribution

• Un coefficient ***nul*** 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.

2) Les coefficients d'aplatissement : En théorie des probabilités et en statistique, le kurtosis (du nom féminin grec ancien « courbure »), aussi traduit par coefficient d'acuité, coefficient d'aplatissement et degré de voussure, est une mesure directe de l'acuité et une mesure indirecte de l'aplatissement de la distribution d'une variable aléatoire réelle. Il existe plusieurs mesures de l'acuité et le kurtosis correspond à la méthode de Pearson.

Étant donnée une variable aléatoire réelle X d'espérance μ et d'écart type σ, on définit son kurtosis non normalisé comme le moment d'ordre quatre de la variable centrée réduite : $\beta_2 = \mathbb{E}[(\frac{X-\mu}{\sigma})^4]$,

lorsque cette espérance existe. On a donc : $\beta_2 = \frac{\mu_4}{\sigma^4}$.

Prenons les coefficients d'asymétrie et d'aplatissement de Pearson ($\beta_1 = \frac{\mu^3}{\sigma^3}$ et $\beta_2 = \frac{\mu^4}{\sigma^4}$), la seule différence avec ceux de Fisher est que le second coefficient n'est pas normalisé, c-à-d $\beta_2 = 3$, pour la loi normale.

On propose les estimateurs

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

La loi conjointe de ces estimateurs est normale bivariée, on écrit

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

La matrice de variance 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 également que la covariance de $\hat{\beta}_1$ et $\hat{\beta}_2$ est nulle.

La forme quadratique associée permet de produire la statistique de Jarque-Bera T qui s'écrit

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

Elle est distribuée asymptotiquement selon une loi du $\mathcal{X}^2$ à 2 degrés de liberté.

α	$\mathcal{X}^2_{1-\alpha}$
0.05	5.99
0.01	9.21

TAB. 2.3 – Exemples des valeurs critiques de test Jarque Bera.

La statistique T prend des valeurs d'autant plus élevées que l'écart entre la distribution empirique et la loi normale est manifeste. La région critique pour un risque α du test est définie par

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

Pour un risque $\alpha = 0.05$, le seuil critique est $\mathcal{X}^2_{0.95} = 5.99$.

Programmation en code R

x=rnorm(100)

jarque.bera.test(x)

Résultat de la commande

Jarque Bera Test

data : x

X-squared = 0.15417, df = 2, p-value = 0.9258

Commentaire : On constate qu'au seuil de risque $\alpha = 5\%$, l'hypothèse de normalité ne peut être rejetter car p-value $> \alpha$.

Conclusion

Le but de notre mémoire est l'étude des techniques statistiques destinées à examiner la compatibilité d'une distribution empirique avec la loi normale. Vérifier la normalité des données continues est une étape cruciale avant la réalisation d'un test d'hypothèse mettant en jeu une ou plusieurs variables continues. 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'hypothèse classiques sont applicables. Si la condition de normalité est violée, il faudra trouver une alternative dite "non paramétrique" au test d'hypothèse à réaliser.

Il existe deux méthodes pour vérifier la normalité :

• La première méthode est la méthode graphique : On peut représenter les données à l'aide de l'histogramme de fréquence et regarder si elles semblent s'ajuster à une distribution normale.

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. On représente les quantiles théoriques en fonction des quantiles observés (Diagramme Q-Q).

• La deuxième méthode est la méthode théorique : Il existe plusieurs tests pour affirmer la normalité d'une distribution. Tous ces tests ont en commun d'avoir comme hypothèse nulle : La distribution empirique suit une loi Gaussienne. Par exemple test de Shapiro-Wilk et test de Kolmogorov-Smirnov.

Bibliographie

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

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

[3] Rahmouni Yasmine, (2019). Test de normalié (mémoire). Université Mohamed Kheider de biskra.

[4] Jean-Jacques Ruch, (2012-2013). STATISTIQUE : TESTS D'HYPOTHESES. Préparation à l'Agrégation Bordeaux.

[5] Djamel Meraghni & Abdelhakim Necir, (2020). Cours de première MASTER. Université Mohamed Khider, Biskra.

[6] NORMALITE DES RESIDUS, http ://d1n7iqsz6ob2ad.cloudfront.net.

[7] Adjengue, L. (2014). Méthodes statistiques : concepts, applications et exercices. Presses internationales Polytechnique.

[8] Pierre Dusart, (2018). Cours de Statistiques inférentielles. Licence 2-S4 SIMASS.

[9] Michel Lejeune, (2010). Statistique La théorie et ses applications. Deuxième édition avec exercices corrigés.

[10] Farida Laoudj Chekraoui, (2016). COURS DE STATISTIQUE MATHEMATIQUE (1) Master Mathématiques appliquées.

[11] B. Desgraupes, (2014-2015). Méthodes Statistiques (L2 Économie) UNIVERSITÉ PARIS OUEST NANTERRE LA DÉFENSE.

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'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 abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous

X	:	Variable aléatoire.
μ	:	La moyenne.
σ^2	:	La variance.
$\mathcal{N}(\mu, \sigma^2)$	:	Loi normale de paramètres μ et σ^2.
F_X	:	Fonction de répartition de v.a X.
F_n	:	Fonction de répartition empirique.
$E(.)$	:	Espérance mathématique.
$V(.)$	:	Variance mathématique.
Φ	:	Fonction de répartition de la loi normale centrée réduite.
H_0	:	Hypothèse nulle.
H_1	:	Hypothèse alternative.
θ	:	Paramètre inconnu.
Θ	:	Ensemble des valeurs de θ.

v.a	:	Variable aléatoire.
tq	:	Tel que
$\overline{X}$	:	Moyenne empirique.
S^2	:	Variance empirique.
S	:	L'écart type
ddl	:	Degré de liberté.
$\mathcal{L}$	:	Convergence en loi.
i.i.d	:	Indépendants identiquements distribués.
Ps	:	Convergence presque sur.
$R.C$	:	Région critique.

Résumé

La distribution la plus utilisée dans l'analyse statistique est la distribution normale, aussi appelée la distribution Gaussienne. Elle est très importante en statistique et en probabilité à cause que la plupart des données en réalité suivent la distribution normale. Dans ce travail, on a présenté deux façons pour vérifier la normalité: la méthode graphique qui consiste à utiliser l'histogramme de fréquence, la boite à moustache, …. La méthode théorique est basée sur : test de Kolmogorov-Smirnov, test de Shapiro-Wilk, test de Lilliefors, test d'Anderson-Darling, ...

Mots clés : Test d'hypothèses, hypothèse, risque, région critique, statistique, test de normalité.

Abstract

The distribution most commonly used in statistical analysis's the normal distribution. It is very important in statistics and probability because in fact most of the data follow a normal distribution. In this work, we have presented two ways to check normality: the graphical method which consists in using histograms, box plot, …. The theoretical method is based on the Kolmogorov-Smirnov test, Shapiro-Wilk test, Lilliefors test, Anderson-Darling test, ….

Key words: Hypothesis test, hypothesis, risk, critical region, statistis, normality test.

ملخص

التوزيع الأكثر استخداما في التحليل الإحصائي هو التوزيع الطبيعي، و يسمى أيضا التوزيع الغاوسي. فهو مهم جدا في الإحصاء و الاحتمالات لأن معظم البيانات في الواقع تتبع التوزيع الطبيعي. في هذا العمل ، قدمنا طريقتين للتحقق من اعتدالية التوزيع : الطريقة البيانية التي تعنى باستخدام المدرج تكراري، مخطط العلبة، .... تعتمد الطريقة النظرية على اختبار كولموغوروف-سميرنوف، اختبار شابيرو-ويلك، اختبار ليليفورس، اختبار اندرسون-دارلينغ، ....

الكلمات المفتاحية: اختبار الفرضية، الفرضية، المخاطر، المنطقة الحرجة، الإحصاءة، اختبار الاعتدالية.