République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE MOHAMED KHIDER, BISKRA

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

DEPARTEMENT DE MATHEMATIQUES

Mémoire présenté par

Mansouri Fouzia

En vue de l'obtention du Diplôme de

MASTER en Mathématiques

Option : **Statistique**

Titre

Tests de Normalité Multivariée

Membres du Comité d'Examen

Pr.	**NECIR Abdelhakim**	UMKB	Président
Pr.	**MERAGHNI Djamel**	UMKB	Encadreur
Dr.	**BENBRIKA Ghozlene**	UMKB	Examinateur

Septembre 2020

Dédicace

Je dédie ce mémoire à

Mes chers parents

Ma mère Salima*, qui a oeuvré pour ma réussite, de par son soutien et ses précieux conseils,*

pour toute son assistance et sa présence dans ma vie.

Mon père Ammar*, qui peut être fier et trouver ici le résultat de longues années de sacrifices*

et de privations pour m'aider à avancer dans la vie.

Mes sœurs Hanane*,* Chahinez*,* Hadjer*,* Ahlame*,* Narimane*.*

Mon cher frère Imade Addine*.*

En leurs souhaitant tout le succès...tout le bonheur.

Mes chères amies Besma, *Oumaima*, *Saoussen*, *Houda, Ahlame.*

Mes professeurs, qui doivent voir dans ce travail la fierté d'un savoir bien acquis.

A tous les étudiants de mathématiques, surtout $2^{éme}$ *master groupe de statistique et tous les étudiants de l'université Mohamed Khider*.

Fouzia Mansouri

REMERCIEMENTS

D'abord je profite de cette occaion pour remercier ***ALLAH*** le tout puissant et miséricordieux

qui m'a donné la force et la patience d'accomplir ce travail.

Je tiens à remercier mon encadreur *Pr. MERAGHNI* Djamel pour la suivi et l'aide qu'il m'a apporté,

et pour ses précieux conseils et aides durant toute la période de préparation de ce mémoire.

Je tiens aussi à remercier l'ensemble des enseignants du département de mathématiques.

Je remercie les membres du jury :

Pr. NECIR Abdelhakim et *Dr. BENBRIKA Ghozlene*.

Enfin, je tiens également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Merci à Tous.

Table des matières

Table des figures

Liste des tableaux

Introduction

En statistique, Les tests de normalité multivariée sont des tests d'hypothèses utilisés pour déterminer si un ensemble d'obsevations multivariées pourrait provenir d'une distribution normale multivariée. Ces tests de normalité prennent une place importante en statistique. Ils permettent de vérifier si des données vectorielles suivent une loi normale multivariée ou non. Pour cela, de nombreuses procédures graphiques sont suggérées. Une possibilité est de vérifier chaque variable séparément pour la normalité univariée.

Dans ce mémoire, on présente dans un premier temps les techniques descriptives. notamment le très populaire graphique, telle que le quantile-quantile plot (Q-Q plot). Dans un second temps, on détaille plusieurs tests statistiques reconnus et implémentés dans la plupart des logiciels de statistique, telle que le logiciel **R**. Enfin, on applique les tests statistiques et l'ajustement graphique par Q-Q plot sur quelques échantillon.

Dans le cadre de ce mémoire, on étudie les moyens permettant de vérifier l'hypothèse de normalité et les méthodes de transformation des observations non normales en observations qui sont à peu près normales. Ce mémoire se compose de deux chapitres :

- **Premier chapitre :** vecteurs aléatoires gaussiens. Ce chapitre est dédié aux vecteurs aléatoires gaussiens et à leurs caractéristiques et propriétés fondamentales.
- **Deuxième chapitre :** tests d'ajustement multinormal. Dans ce chapitre, après un bref rappel sur les différents tests d'ajustement, on présente, avec plus de détails quelques méthodes de valider la normalité multivariée. Enfin, on traite des exemples d'application de ces tests à l'aide du logiciel d'analyse statistique **R**.

Chapitre 1

Vecteurs aléatoires gaussiens

Dans ce chapitre, on étudie les caractéristiques des vecteurs aléatoires (V.a) gaussiens en particulier. Ces derniers représentent une généralisation des variables aléatoires (v.a) réelles normales.

1.1 Variables aléatoires

Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé où Ω est un ensemble fondamontal lié à une expérience aléatoire, $\mathcal{F}$ est une tribu de Ω et P est une mesure de probabilité définie sur $\mathcal{F}$ par

$$\begin{array}{rccc} P: & (\Omega, \mathcal{F}) & \rightarrow & [0,1] \\ & A & \rightarrow & P(A) \end{array}.$$

Définition 1.1.1 *On appelle v.a sur $(\Omega, \mathcal{F}, P)$ toute application mesurable de Ω dans $\mathbb{R}$. On la note généralement par X.*

$$\begin{array}{rccc} X: & \Omega & \rightarrow & \mathbb{R} \\ & \omega & \mapsto & x = X(\omega) \end{array}.$$

Il existe deux types de v.a Les v.a discrètes et les v.a continues. On dit que X est discrète

si les valeurs qu'elle prend appartiennent à un ensembre dénombrable, c-a-d l'ensemble $X(\Omega)$ *est un sous ensemble dénombrable de* $\mathbb{R}$. *Et on dit que* X *est continue, si l'ensemble des ses valeurs qu'elle peut prendre appartiennent à un ensemble nom dénombrable inclu dans* $\mathbb{R}$.

Exemple 1.1.1 *On lance deux dés distincts. Alors* $\Omega = \{(1,1),(1,2),...,(6,5),(6,6)\}$. *On désigne par* X *la somme des deux numéros obtenus. Alors l'ensemble des valeurs possibles,* $X(\Omega) = \{2,3,4,...,12\}$, *est dénombrable. Donc* X *est une v.a discrète.*

Exemple 1.1.2 X *mesure de taille ou de poids d'un élève est une v.a continue.*

1.1.1 Loi d'une variable aléatoire

Fonction de répartition

Définition 1.1.2 *On appelle fonction de répartition d'une v.a* X *l'application*

$$\begin{array}{rccc} F: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & F(x) := P(X \leq x) \end{array}$$

C'est une fonction comprise entre 0 et 1, continue à droite et croissante sur $\mathbb{R}$. Son inverse généralisé est appelé fonction des quantiles (voir [1]).

Cas discret

Définition 1.1.3 *Soit* X *une v.a discrète, alors sa loi est déterminée par la probabilité* P_X *définie par*

$$p_x = P_X(x) := P(X = x) = P\left(X^{-1}(\{x\})\right) = P(\{\omega \in \Omega / X(\omega) = x\}).$$

Dans ce cas, la fonction de répartition devient

$$F(x) = \sum_{s \leq x} P_X(s), \ x \in \mathbb{R}.$$

Cas continu

La loi de probabilité d'une v.a continue est définie par ce que l'on appelle fonction de densité.

Définition 1.1.4 *On appelle fonction de densité toute application f définie sur $\mathbb{R}$ satisfaisant :*

1. $\forall x \in \mathbb{R},\ f(x) \geq 0$;
2. $\int_{-\infty}^{+\infty} f(x)dx = 1.$

La fonction de répartition est alors donnée par

$$F(x) = \int_{-\infty}^{x} f_X(s)ds,\ x \in \mathbb{R}.$$

En d'autres termes, f est la dérivée de F.

Remarque 1.1.1 *Pour a et b dans $\mathbb{R}$, on a*

$$P(a \leq X \leq b) = \int_a^b f(s)ds = F(b) - F(a).$$

1.1.2 Moments d'une variable aléatoire

Une loi de probabilité peut être caractérisée par certaines valeurs typiques associées aux notions de valeur centrale, de dispersion et de forme de la distribution.

Définition 1.1.5 *On appelle moment d'ordre $k \geq 1$, d'une v.a X la quantité*

$$\mu^{(k)} = E[X^k] := \begin{cases} \sum_x x^k P_X(x) & si \quad X \text{ est disrète,} \\ \int_{-\infty}^{+\infty} x^k f(x)dx & si \quad X \text{ est continue.} \end{cases}$$

Espérance mathématique-Variance

L'espérance d'une v.a est égale au moment (particulier) d'ordre 1. Elle correspond à la moyenne des valeurs possibles de X pondérées par les probabilités associées à ces valeurs. C'est un paramétre de position qui réprésente l'équivalent théorique de la moyenne arithmétique (empirique) [1]. On la note généralement par μ :

$$\mu = \mu^{(1)} = E[X] \tag{1.1}$$

La variance d'une varaible aléatoire est un paramètre de dispersion [8], défini par

$$\sigma^2 = Var(X) := E[X - E[X]]^2 = E[X^2] - E[X]^2.$$

Sa racine carrée σ est appelée écart type de X.

1.1.3 Fonction caractéristique

Définition 1.1.6 *La fonction caractéristique d'une v.a est l'éspérance mathématique de la variable complexe* $\exp(itX)$ *[9], c-à-d*

$$\phi_X(t) := E[\exp(itX)],\ t \in \mathbb{R}. \tag{1.2}$$

Proposition 1.1.1 *Si deux v.a X et Y sont reliées par la relation $Y = \alpha X$, alors on a*

$$\phi_Y(t) = \phi_X(\alpha t)$$

Proof. D'après (1.2), on a

$$\phi_Y(t) = E[exp(itY)].$$

On remplace Y par αX et on trouve

$$\phi_Y(t) = E[exp(\alpha itX) = E[\exp(i(\alpha t)X) = \phi_X(\alpha t).$$

■

Proposition 1.1.2 *La fonction caractéristique dela somme de v.a indépendantes $X_1, ...X_p$ est égale au produit des fonctions caractéristiques individuelles [3].*

$$\phi_{X_1+...X_p}(t) = \prod_{k=1}^{p} \phi_{X_k}(t).$$

Proof. Soient $X_1, ...X_p$, des v.a indépendantes de fonctions caractéristiques $\phi_{X_1}, ..., \phi_{X_p}$, et de somme $Y = \sum_{k=1}^{p} X_k$.La fonction caractéristique de Y s'écrit

$$\phi_Y(t) = E[e^{itY}] = E[e^{it\sum_{k=1}^{p} X_k}] = E[\prod_{k=1}^{p} e^{itX_k}].$$

Puisque l'espérance du produit égale au produit des espérances, alors ona

$$\phi_Y(t) = \prod_{k=1}^{p} E[e^{itX_k}] = \prod_{k=1}^{p} \phi_{X_k}(t).$$

■

1.1.4 Loi normale unidimensionnelle

Définition 1.1.7 *On appelle loi normale de paramétre $\mu \in \mathbb{R}$ et $\sigma \geq 0$ la loi d'une variable aléatoire continue X prenant toutes les valeurs réelles, de densité de probabilité la fonction définie pour tout $x \in \mathbb{R}$ par*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2].$$

La fonction de répartition est

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2]dt, \ x \in \mathbb{R}.$$

Loi et symbole	Probabilité	$E[X]$	$Var(X)$	Fonction caractéristique
Bernouilli $\mathcal{B}(p)$	$p^k(1-p)^{1-k}\,\amalg_{\{0,1\}}(k)$	p	$p(1-p)$	$1-p+pe^{it}$
Binomiale $\mathcal{B}(n,p)$	$C_n^k p^k(1-p)^{n-k}\,\amalg_{\{0,\dots,n\}}(k)$	np	$np(1-p)$	$1-p+pe^{it}$
Poisson $\mathcal{P}(\lambda)$	$e^{(-\lambda)}\frac{\lambda^k}{k\uparrow}\,\amalg_{\mathbb{N}}(k)$	λ	λ	$e^{\lambda(e^{it}-1)}$
Géométrique $\mathcal{G}(p)$	$p(1-p)^{k-1}\,\amalg_{\mathbb{N}^*}(k)$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$	$\frac{pe^{it}}{1-(1-p)e^{it}}$

TAB. 1.1 – Quelques lois de probabilité discrètes usuelles

Son espérance et variance sont respectivement égales à

$$E[X]=\mu \text{ et } Var(X)=\sigma^2.$$

La loi normale standard

Définition 1.1.8 *Une v.a Z est dite gaussienne centrée réduite ou de loi normale standard, ce qu'on note $X\sim\mathcal{N}(0,1)$, si sa loi admet pour densité*

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

par rapport à le mesure de Lebesgue sur $\mathbb{R}$.

1.1.5 Lois de probabilités usuelles

Les caractéristiques de quelques lois de probabilité usuelles sont résumées dans les tableaux 1.1 (cas discret) et 1.2 (cas continu), avec $n\in\mathbb{N}^*$, $p\in]0,1[$, $\lambda\in\mathbb{R}_+^*$, $a<b\in\mathbb{R}$, $\mu\in\mathbb{R}$, et $\sigma\in\mathbb{R}_+^*$.

loi et symbole	densité	$E[X]$	$Var(X)$	fonction caractéristique
uniforme $\mathcal{U}[a,b]$	$\frac{1}{b-a}\amalg_{[a.b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
normale $\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{[-\frac{1}{2}((x-\mu)/\sigma)^2]}$	μ	σ^2	$e^{it\mu-\sigma^2t^2/2}$
exponentielle $\mathcal{E}(\lambda)$	$\lambda e^{(-\lambda x)}\amalg_{\mathbb{R}_+}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-it}$

TAB. 1.2 – Quelques lois de probabilité continues usuelles

1.1.6 Théorème central limite

Le théorème central limite établit la convergence en loi de la somme d'une suite de v.a indépendantes vers la loi normale. Intuitivement, ce résultat affirme que toute somme de v.a iid tend vers une v.a gaussienne [3].

Soit (X_n) une suite de v.a définies sur le même espace de probabilité, suivant la même loi avec $\mu = E(X_1) < \infty$ et $\sigma^2 = Var(X_1) < \infty$. De plus, on suppose qu'elles sont indépendantes. Si $S_n = X_1 + ... + X_n$, alors l'espérance de S_n est $n\mu$ et sa variance $n\sigma^2$ et $S_n - E[S_n]/\sqrt{var(s_n)}$ converge en loi vers une v.a normale centrée réduite, c-à-d

$$\frac{S_n - n\mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1), \text{ quand } n \to \infty.$$

1.2 Vecteurs aléatoires

1.2.1 Définitions et propriétés fondamentales

Définition 1.2.1 *Un V.a X est une application de $(\Omega, \mathcal{F}, P)$ dans un espace vectoriel réel, en général $\mathbb{R}^p$ muni de sa tribu borélienne. En pratique $\mathbb{R}^p$ est muni de sa base canonique et on écrira $X = (X_1, ..., X_p)^t$. Les v.a $X_1, ..., X_p$ sont dites composantes de X.*

Espérance-matrice de covariance

Définition 1.2.2 *L'espérance (moyenne, moment d'ordre 1) de X est le vecteur de $\mathbb{R}^p$*

égale à

$$E[X] := (E[X_1], ..., E[X_p])^t.$$

On la note généralement par $\mu = (\mu_1, ..., \mu_p)^t$ *avec* $\mu_j := E([X_j], j = 1, ..., p,$ *comme définie par* (1.1) .

Remarque 1.2.1 *On dit que le V.a* X *est centré si* $E[X]$ *est le vecteur nul dans* $\mathbb{R}^p$.

Définition 1.2.3 *On appelle matrice de covariance de* X *la matrice* Σ, *définie par*

$$\Sigma := E[(X - \mu)(X - \mu)^t] = E[XX^t] - \mu\mu^t.$$

C'est une matrice carrée de taille p, *symétrique dont les coeffcients sont*

$$\sigma_{ij} = Cov(X_i; X_j) := E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i\mu_j, \ i, j = 1, ..., p.$$

On a donc

$$\Sigma = \begin{bmatrix} \sigma^2_{1} & \sigma_{12} & \cdots & \cdots & \sigma^2_{1p} \\ \sigma^2_{21} & \sigma^2_{2} & \ddots & \cdots & \sigma^2_{2p} \\ \vdots & \vdots & . & \ddots & \vdots \\ \vdots & \vdots & . & . & \vdots \\ \sigma^2_{p1} & \sigma^2_{p2} & \cdots & \cdots & \sigma^2_{p} \end{bmatrix},$$

où $\sigma^2_j = \sigma_{jj} = Var\,(X_j)$.

Il est à noter que la matrice de covariance joue un rôle de grande importance en statistique multivariée. En effet, elle permet de résumer les dépendances linéaires entre tous les couples de composantes d'un vecteur aléatoire.

Proposition 1.2.1 *Si les variables coordonnées* $X_1, ..., X_p$ *sont indépendantes, la matrice de covariance du vecteur* X *est diagonale :* $Cov(X_i; X_j) = 0$, $i \neq j$.

Proposition 1.2.2 *La matrice de covariance d'un V.a est une matrice symétrique semi-définie positive.*

Définition 1.2.4 *On appelle matrice de corrélation de X la matrice carrée symétrique R, de taille p, dont les coeffcients sont donnés, pour $i, j = 1, ..., p$, par $r_{ij} = Cor(X_i; X_j) := \sigma_{ij}/(\sigma_i \sigma_j)$:*

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & \cdots & r_{1p} \\ r_{21} & 1 & & & r_{2p} \\ \vdots & \vdots & . & & \vdots \\ \vdots & \vdots & & . & \vdots \\ r_{p1} & r_{p2} & ... & ... & 1 \end{bmatrix}$$

Remarque 1.2.2 *Si les variables X_j sont réduites $\left(\sigma_j^2 = 1\right)$, alors Σ s'identifie avec R.*

Fonction de répartition d'un V.a

Tout V.a admet une fonction de répartition, notée F ou F_X, qui est une application de $\mathbb{R}^p$ dans $\mathbb{R}$ définie par

$$F(x) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_p \leq x_p),\ x = (x_1, x_2, ..., x_p) \in \mathbb{R}^p.$$

Fonction caractéristique d'un V.a

Définition 1.2.5 *On appelle fonction caractéristique du V.a $X = (X_1, ..., X_p)^t$ la fonction a valeurs complexes définie par*

$$\phi_X(u) := E[e^{i\langle u, X\rangle}] = E[e^{iu^t X}] = E[e^{i\sum_{j=1}^p u_j x_j}],\ u = (u_1, ...u_p)^t \in \mathbb{R}^p.$$

Lorsque le vecteur X admet une densité f_x, on a

$$\phi_X(u) = \int_{\mathbb{R}^p} e^{i\sum_{j=1}^p u_j x_j} f_X(x_1, ...x_p) dx_1 ... dx_p],\ u = (u_1, ...u_p)^t \in \mathbb{R}^p.$$

La fonction caractéristique est donc la transformée de Fourier de la densitée f_X. La densité f_x s'exprime en fonction de ϕ_X a l'aide de la transformée inverse

$$\forall x \in \mathbb{R}^p,\ f_X(x) = \frac{1}{(2\pi)^p} \int_{\mathbb{R}^p} e^{-i\sum_{j=1}^p u_j x_j} \phi_X(u_1, ...u_p) du_1 ... du_p.$$

Comme son nom l'indique, la fonction caractéristique permet de caractériser la loi du vecteur.

Proposition 1.2.3 ***(unicité)***. *Soient X et Y deux v.a. On note ϕ_X et ϕ_Y leurs fonctions caractéristiques. Si $X = Y$, alors elles admettent la même loi $\phi_X = \phi_Y$.*

Proposition 1.2.4 *Soit X un V.a à valeurs dans $\mathbb{R}^p$. Une probabilité μ sur $(\mathbb{R}^p, B(\mathbb{R}^p))$ est la loi de X si et seulement si, pour toute fonction φ de $\mathbb{R}^p$ dans $\mathbb{R}$ mesurable, positive, on a*

$$E(\varphi(x)) = \int_{\mathbb{R}^n} \varphi(t) d\mu(t).$$

Remarque 1.2.3 *L'espérance d'un produit de fonctions de v.a indépendantes est égale au produit des espérances. Donc, on en déduit que*

$$E[exp(iu^t X)] = E[\exp(iu_1 X_1)]...E[\exp(iu_p X_p)].$$

1.2.2 Vecteurs aléatoires à densité

Densité d'une distribution multivariée

Définition 1.2.6 *La loi du V.a X est une mesure de probabilité sur $\mathbb{R}^p$ muni de sa tribu de Borel $B(\mathbb{R}^p)$. Elle est notée P_X et se définit par*

$$\forall B \in B(\mathbb{R}^p),\ P_X(B) = P(X \in B).$$

On dit que le vecteur X admet une densité si la loi P_X admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$. Cette densité est alors notée par f ou f_X. Dans ce cas, on

écrit

$$\forall X \in \mathbb{R}^p,\ F_X(x) = P(X \in B) = \int_{-\infty}^{x_1} .... \int_{-\infty}^{x_p} f(x_1, ..., x_p) dx_1 ... dx_p.$$

De plus, dans le cas où f existe et est continue, on a

$$f(x) = \frac{\partial^p F}{\partial x_1 \partial x_2 ... \partial x_p}(x_1, x_2, ...x_p),\ x = (x_1, x_2, ...x_p) \in \mathbb{R}^p.$$

Ceci généralise la relation bien connue dans le cas unidimensionnel.

Remarque 1.2.4 *On rappelle que les variables $X_1, ..., X_p$ sont indépendantes si et seulement si*

$$\forall B_1, ..., B_p \in B(\mathbb{R}),\ P(X_1 \in B_1, ..., X_p \in B_p) = P(X_1 \in B_1) ... P(X_p \in B_p).$$

Densité marginale

Lorsque $X = (X_1, ..., X_p)^t$ est continu de densité f_X, chacune de ses composantes X_j $(j = 1, ..., p)$ est continue avec densité f_j définie sur $\mathbb{R}$ par

$$f_j(t) = \int_{\mathbb{R}^{p-1}} f_X(u_1, ..., t, ...u_p) du_1 ... du_{j-1} du_{j+1} ... du_p.$$

1.2.3 Indépendance, dépendance

Indépendance de deux v.a.

On considère ici deux v.a X et Y . On dit que X et Y sont indépendantes si la loi du couple (X, Y) est le produit des lois de X et Y, ce qui revient à dire que $\forall x, y \in \mathbb{R}$

$$\begin{aligned} F_{(X,Y)}(x, y) &= P_{(X,Y)}\left(]-\infty, x] \times]-\infty, y]\right) \\ &= P_X\left(]-\infty, x]\right) \times P_Y(]-\infty, y]) = F_X(x) F_Y(y). \end{aligned}$$

Propriété 1.2.1 *Si X et Y sont deux v.a indépendantes alors, pour toutes fonctions continues $f : \mathbb{R} \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$ telles que $E[|g(X)|] < +\infty$ et $E[|h(Y)|] < +\infty$, on a*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Remarque 1.2.5 *L'indépendance implique la non-corrélation.*

En effet, pour X et Y indépendantes on a $Cov[X, Y] = E[g(X)h(Y)] = E[g(X)]E[h(Y)]$, où $g(x) = (x - E[X])$ et $h(y) = (y - E[Y])$. Or, il est clair que $E[g(X)] = E[h(Y)] = 0$.

Fonction caractéristique d'un V.a de composantes indépendantes

Proposition 1.2.5 *Les composantes $X_1, ..., X_p$ du V.a X sont indépendantes si et seulement si la fonction caractéristique du vecteur X est égale au produit des fonctions caractéristiques de ses composantes.*

$$\phi_X(u) = \prod_{j=1}^{p} \phi_{Xj}(u_j), \ u = (u_1, ...u_p)^t \in \mathbb{R}^p.$$

Proof. [$\Longrightarrow$] Supposons que les composantes de X sont indépendantes. Pour tout $u = (u_1, ...u_p)^t \in \mathbb{R}^p$ on a

$$\phi_X(u) = E[e^{i\sum_{j=1}^{p} u_j x_j}] = E\left[\prod_{j=1}^{p} e^{iu_j x_j}\right] = \prod_{j=1}^{p} E[e^{iu_j x_j}] = \prod_{j=1}^{p} \phi_{X_j}(u_j).$$

[$\Longleftarrow$] Réciproquement pour tout $u = (u_1, ...u_p)^t \in \mathbb{R}^p$, on a

$$\phi_X(u) = \prod_{j=1}^{p} \phi_{X_j}(u_j),$$

c-à-d

$$\int e^{i\sum_{j=1}^{p} u_j x_j} dP_{X_1,...,X_j}(x_1, ..., x_p) = \prod_{j=1}^{p} \int e^{iu_j x_j} dP_{X_j}(x_j) = \int e^{i\sum_{j=1}^{p} u_j x_j} \prod_{j=1}^{p} dP_{X_j}(x_j).$$

L'unicité de la fonction caractéristique implique que la loi de$X$ est une mesure produit. D'où l'indépendance des composantes. ■

Densité d'un V.a.c de composantes indépendantes

Lorsque X est continu et admet une densité f_X,cette dernière se décompose en produit des densités marginales, c-à-d

$$\forall x = (x_1, ..., x_p) \in \mathbb{R}^p,\ f_X(x) = \prod_{j=1}^{p} f_{X_j}(x_j).$$

1.3 Vecteur aléatoires gaussiens

Définition 1.3.1 *Un V.a $X = (X_{1,}..., X_p)^t$ est gaussien si et seulement si toute combinaison linéaire*

$$\langle \lambda, X \rangle_{\mathbb{R}^p} = \lambda^t X = \sum_{j=1}^{p} \lambda_j X_j,\ \lambda = (\lambda_1, ..., \lambda_p)^t \in \mathbb{R}^p,$$

de ses composantes est une v.a gaussienne.

Remarque 1.3.1 *Si X est un V.a gaussien, alors chacune de ses composantes est une v.a gaussienne. La réciproque n'est pas forcément vraie.*

Si $X = (X_{1,}..., X_p)^t$ est un vecteur gaussien, on définit son espérance (vecteur moyenne) $E[X]$ par

$$\mu = E[X] := (E[X_1], ..., E[X_p])^t,$$

et sa matrice de covariance $Var(X)$ par

$$\Sigma = Var(X) := E[(X - \mu))(X - \mu)^t] = E[XX^t] - \mu\mu^t.$$

On note $X \sim \mathcal{N}_p(\mu, \Sigma)$.

1.3.1 Propriétés des vecteurs gaussiens

Proposition 1.3.1 *Si* $X \sim \mathcal{N}_p(\mu, \Sigma)$, *alors sa fonction caractéristique est*

$$\forall u \in \mathbb{R}^p,\ \phi_X(u) = \exp(iu^t\mu - \frac{1}{2}u^t\Sigma u).$$

Remarque 1.3.2 (indépendance) *Une propriété importante des vecteurs gaussiens est que l'indépendance des composantes est équivalente à la nullité des covariances, c-à-d Si* X *est un vecteur gaussien, alors pour tout couple* (i,j), $i \neq j$, X_i *et* X_j *sont indépendantes si et seulement si leur covariance est nulle.*

Proposition 1.3.2 (linéarité) *Soient* $X = (X_1,...,X_p)^t \sim \mathcal{N}_p(\mu, \Sigma)$, A *une matrice* $q \times p$ *(*q *lignes et* p *colonnes) et* b *un vecteur de* $\mathbb{R}^p$. *Alors* $Y := AX + b$ *est un vecteur gaussien a valeurs dans* $\mathbb{R}^q$ *avec*

$$E[Y] = A\mu + b,\ \text{et } \Sigma_Y = A\Sigma A^t,$$

Σ_Y *désignant la matrice de covariance de* Y. *En d'autres termes, on a* $Y \sim \mathcal{N}_q(A\mu + b, A\Sigma A^t)$.

Proposition 1.3.3 (densité) *Si* $X \sim \mathcal{N}_p(\mu, \Sigma)$, *alors la loi de* X *admet une densité* $f_{\mu,\Sigma}$ *par rapport à la mesure de Lebesgue sur* $\mathbb{R}^p$ *si et seulement si* Σ *est inversible. Dans ce cas la densité est*

$$f_{m,\Sigma}(x) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right).$$

Cas particulier $p = 2$

Soit $X = (X_1, X_2)^t \sim \mathcal{N}_2(\mu, \Sigma)$ avec $\mu = (\mu_1, \mu_2)^t$ et

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

où ρ est le coefficient de corrélation linéaire entre X_1 et X_2. Alors $\det \Sigma = (1-\rho^2)(\sigma_1\sigma_2)^2$ et

$$\Sigma^{-1} = \frac{1}{(1-\rho^2)(\sigma_1\sigma_2)^2}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}.$$

Dans ce cas, la densité de probabilité du vecteur s'écrit

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

Remarque 1.3.3 *La loi d'un vecteur gaussien est entièrement déterminée par sa moyenne μ et sa matrice de covariance Σ.*

Proposition 1.3.4 *Soient $X_1,...,X_p$ des v.a indépendantes de loi $\mathcal{N}(0,1)$. Alors le V.a $X = (X_1,...,X_p)^t$ est gaussien.*

Proof. On considère une combinaison linéaire Y des X_j. Soient $a_1....a_p$ des scalaires réels et $Y = \sum_{j=1}^p a_j X_j$. La fonction caractéristique de Y est

$$\phi_Y(t) = \prod_{j=1}^p \phi_{X_j}(a_j t) = \prod_{j=1}^p \exp(-\frac{1}{2}a_j^2 t^2) = \exp(-\frac{t^2}{2}\sum_{j=1}^p a_j^2).$$

Donc, la variable Y suit la loi $\mathcal{N}(0, \sum_{j=1}^p a_j^2)$. ■

1.3.2 Lois conditionnelles

On divise X en deux sous-vecteurs X_1 et X_2, à k et $p-k$ composantes, d'espérances μ_1 et μ_2.respectivement.

$$X = \begin{bmatrix} X_1 \\ \hline X_2 \end{bmatrix}$$

La matrice de variance-covariance se partitionne en 4 blocs

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Si on cherche la loi du vecteur X_1 conditionné par X_2 on a les résultats suivants :

Proposition 1.3.5 *La loi de X_1/X_2 est une loi multinormale à k dimensions :*

- d'espérance $E[X_1/X_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$;

- de matrice covariance $\Sigma_{11/2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

On constate donc que la régression de X_1 en X_2 est linéaire. Les termes de $\Sigma_{11/2}$ s'appellent les covariances partielles $cov(i, j/2)$, desquelles on déduit les corrélations partielles

$$r_{ij/2} = \frac{cov(i, j/2)}{\sigma_{ii/2}\sigma_{jj/2}}.$$

1.3.3 Formes quadratique

Proposition 1.3.6 *Soit un V.a $X \sim \mathcal{N}_p(\mu, \Sigma)$, alors la forme quadratique*

$$D^2 := (X - \mu)^t \Sigma^{-1} (X - \mu) \sim \mathcal{X}_p^2. \tag{1.3}$$

Ce résultat est démontré dans [12], page 95.

1.3.4 Echantillon d'un V.a gaussien

Un échantillon de taille n d'un V.a $X \sim \mathcal{N}_p(\mu, \Sigma)$ est une suite $\left(X^{(1)}, ..., X^{(n)}\right)$ de n V.a iid de même loi que X. On a

$$X^{(i)} = \left(X_1^{(i)}, ..., X_p^{(i)}\right)^t \sim \mathcal{N}_p(\mu, \Sigma), \ i = 1, ..., n,$$

Les moyenne et matrice de covariance empiriques sont respectivement définies par

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X^{(i)} \text{ et } S := \frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)} - \bar{X}\right)\left(X^{(i)} - \bar{X}\right)^{t}.$$

Un résultat très important en inférence statistique concerne la distribution asymptotique de la moyenne empirique $\bar{X}$. Connu sous le nom de théorème central limite multivarié, il généralise le résultat du cas unidimensionnel (voir [10], page 176).

Proposition 1.3.7 (théorème central limite multivarié) *Soient $X^{(1)}, ..., X^{(n)}$ n observations indépendantes d'un V.a X (quelconque) d'espérance μ et de matrice de covariance Σ. Alors, pour n et $(n-p)$ grands,on a*

$$\sqrt{n}\left(\bar{X} - \mu\right) \rightsquigarrow \mathcal{N}_p\left(0, \Sigma\right).$$

Un deuxième résultat très utile en analyse statistique multidimensionnelle peut être obtenu en remplaçant, dans (1.3), μ et Σ par leurs estimations respectives $\bar{X}$ et S (voir [10], page 183)

Proposition 1.3.8 *Soit $\left(X^{(1)}, ..., X^{(n)}\right)$ un échantillon d'un V.a $X \sim \mathcal{N}_p\left(\mu, \Sigma\right)$. Alors, pour n et $(n-p)$ grands,on a l'approximation suivante :*

$$\left(X - \bar{X}\right)^{t} S^{-1}\left(X - \bar{X}\right) \rightsquigarrow \mathcal{X}_p^2. \tag{1.4}$$

Chapitre 2

Tests d'ajustement multinormal

De nombreux tests et procédures graphiques ont été suggérés pour évaluer si un ensemble de données provient d'une population normale multivariée. Une première possibilité est de vérifier chaque composante séparément pour la normalité univariée. Il s'agit donc de tester les hypothèses suivantes :

$$\begin{cases} H_0 : & \text{la composante } X_j \text{ est normale} \\ H_1 : & \text{la composante } X_j \text{ n'est pas normale} \end{cases}$$

Le but est d'avoir une idée sur la distribution adéquate à ces données. L'élément le plus pertinent à la modélisation est le quantile-quantile plot (Q-Q plot).

2.1 Tests d'ajustement

2.1.1 Test de Kolmogorov-smirnov

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

Statistique de test

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

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

où F_0 est la fonction de répartition théorique et

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \amalg_{(X_i\leq x)},$$

désigne la fonction de répartition empirique définie par l'échantillon $(X_1, ...X_n)$.
Pour calculer les valeurs de la statistique D_n , il suffit d'évaluer la diférence entre F_n et F_0 aux points $x_{(i)}$ comme l'indique la proposition suivante.

Proposition 2.1.1 *La statistique de Kolmogorov-Smirnov s'écrit comme suit :*

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

Principe du test

On calcule la distance entre F_n et F_0 en utilisant la relation 2.1 puis on dècide du rejet ou non du modèle proposé. L'exécution du test de Kolmogorov-Smirnov est donnée par les étapes suivantes :

1. classer les valeurs observées par ordre croissant ;

2. calculer, pour $i = 1$, les valeurs des écarts $\left[i/n - F_0(x_{(i)})\right]$ et $\left[F_0(x_{(i)}) - (i-1)/n\right]$;

3. prendre le plus grand des deux écarts ;

4. répéter les étapes 2 et 3 pour $i = 2, ..., n$;

5. la valeur de D_n est égale au maximum entre le plus grand écart et 0.

La région critique du test est de la forme $\{D_n > D_{crit}\}$, où D_{crit} est une certaine valeur critique vérifiant $P(D_n > D_{crit}/H_0 \text{ est vraie}) = \alpha$, $0 \leq \alpha \leq 1$. On conclut le test en acceptant, au seuil de signification α, l'hypothèse H_0 si la distance D_n calculée est

inférieure à D_{crit}.

La valeur critique D_{crit} est lue à partir la table de Kolomogorov-Sminrov.(voir, par exemple,[12], pages $585-586$). Pour les petites tailles d'échantillons, il y a dans [4], pages $113-114$, un exemple de calcul de D_{crit} (pour $n=2$) à partir de la loi exacte de D_n. Pour $n \geq 50$, les valeurs critiques sont données, selon quelques valeurs de α dans le tableau 2.1.

2.1.2 Test de Cramer-von Mises

Le test est développé par Harald Cramer et Richard E. von Mises $(1928-1930)$.

Statistique du test

Ce test est basé sur la différence quadratique entre la fonction de répartition empirique et la fonction de répartition théorique. La statistique du test d'ajustement de Cramer-von Mises $\bar{w}_n^2$ est définie par

$$\bar{w}_n^2 := n \int_{-\infty}^{+\infty} \left(F_n(x) - F_0(x)\right)^2 dF_0(x).$$

En pratique, son calcul est simplifié comme l'indique la proposition suivante.

Proposition 2.1.2

$$\bar{w}_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F_0(x_i) \right)^2. \tag{2.2}$$

Principe du test

On calcule la distance entre F_n et F_0 en utilisant la relation 2.2, puis on décide du rejet ou non du modèle proposé. L'exécution du test de Cramer-von Mises est donnée par les étapes suivantes :

1. classer les valeurs observées par ordre croissant ;

2. utiliser la fonction de répartition de la loi pour obtenir les valeurs de $F_0(x_i)$, pour $i = 1, ..., n$;

3. calculer $\sum_{i=1}^{n} \left((2i-1)/2n - F_0(x_i) \right)^2$ puis la valeur de la statistique $\bar{w}_n^2$.

On rejette l'hypothèse H_0 si cette dernière est supérieure à une certaine valeur critique n'ayant qu'une probabilité α d'être dépassée, sous l'hypothèse H_0. Il existe une table statistique, connue sous le nom de table de Cramer-von Mises, dans laquelle sont résumées les valeurs critiques pour les niveaux de signification usuels avec différentes tailles d'échantillons (voir, par exemple,[12], page 584). Pour $n \geq 50$ les valeurs critiques sont données, selon quelques valeurs de α, dans le tableau 2.1.

2.1.3 Test d'Anderson-Darling

Construit en 1954 par Theordore W. Anderson et Donald A. Darling dans une première version, puis généralisé par Michael A. Stephens en 1974. Il s'agit d'une modification du test de Cramer-von Mises, il donne plus importance aux queues de distribution.

Statistique du test

La statistique du test d'Anderson-Darling $\not{A}_n^2$ est définie par

$$\not{A}_n^2 := n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x)$$

Remarque 2.1.1 *Une simplification de cette statistique est donnée par*

$$A_n^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)[\log(F_0(x_i) + \log(1 - F_0(x_{n-i+1})]\right) - n,$$

ou

$$A_n^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)\log(F_0(x_i) + (2n+1-2i)\log(1 - F_0(x_i))\right) - n. \tag{2.3}$$

Principe du test

On calcule la distance entre F_n et F_0 en utilisant la relation 2.3 puis on décide du rejet ou non du modèle proposé. L'exécution du test d'Anderson-Darling est donnée par les étapes suivantes :

n	Kolmogorov-Smirnov				Cramer-von Mises				Anderson-Darling			
	niveau de signification											
	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
50	0.15	0.16	0.17	0.19	0.24	0.28	0.34	0.45	1.43	1.62	1.90	2.42
100	0.11	0.11	0.12	0.14	0.24	0.29	0.36	0.47	1.39	1.60	1.91	2.41
200	0.08	0.08	0.09	0.10	0.24	0.24	0.28	0.45	1.39	1.59	1.92	1.49
500	0.05	0.05	0.05	0.06	0.23	0.28	0.33	0.44	1.41	1.61	1.93	2.50
800	0.04	0.04	0.04	0.05	0.24	0.29	0.35	0.47	1.41	1.62	1.90	2.40
1000	0.03	0.04	0.04	0.04	0.25	0.29	0.35	0.45	1.42	1.64	1.95	2.51
2000	0.02	0.03	0.03	0.03	0.24	0.29	0.35	0.48	1.39	1.59	1.91	2.44

TAB. 2.1 – Quelques valeurs critiques de Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling (source [14])

1. ordonner les observations de manière croissante ;

2. obtenir les fréquences théoriques $F_0(x_i)$; puis déduire $\log(F_0(x_i)$ et $\log(1-F_0(x_i))$; puis la valeur de la statistique A_n^2;

3. calculer $\sum_{i=1}^{n}(2i-1)\log(F_0(x_i)+(2n+1-2i)\log(1-F_0(x_i))$, puis la valeur de la statistique A_n^2;

4. Pour faire la décision, on compare la valeur A_n^2 avec une certaine valeur critique A_{crit}^2 au certain seuil de signification α, l'hypothèse H_0 est rejetée lorsque la statistique A_n^2 prend des valeurs trop élevées, c-àd lorsque $A_n^2 > A_{crit}^2$.

Les valeurs critiques de A_n^2 sont tabulées (voir, par exemple, [2], page 112 et [14]). Pour $n \geq 50$; les valeurs A_{crit}^2 sont données, selon quelques valeurs de α dans le tableau 2.1.

2.1.4 Test de normalité de Lilliefors

Les tests précédents sont des tests généraux qui s'appliquent à n'importe quelle distribution F_0 de l'hypothèse nulle. Lorsque cette dernière est la loi normale, on parle de test de normalité. Il s'agit donc de vérifier l'ajustement d'un ensemble d'observations à un modèle Gaussien. Les hypothèses suivantes à tester sont

$$\begin{cases} H_0 : & \text{les données suivent une loi normale,} \\ H_1 : & \text{les données ne suivent pas une loi normale.} \end{cases}$$

Pour cela, il existe plusieurs procédures, parmi lesquelles le test de Lilliefors qui a été introduit en 1967 par Hubert Lilliefors. C'est une approche non paramétrique visant à tester si une variable continue X suit une loi normale de paramètres μ et σ^2 inconnus qu'on estime par leurs contre parties empiriques $\bar{x}$ et s_n^2 respectivement. La statistique de Lilliefors L est définie par

$$L := \max\left\{\max_{1\leq i\leq n}\left[\frac{i}{n} - \Phi(z_{(i)})\right], \max_{1\leq i\leq n}\left[\frac{i-1}{n} - \Phi(z_{(i)})\right], 0\right\}, \tag{2.4}$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite et $z_{(i)}$ la valeur ordonnée de $z_i := (x_i - \overline{x})/s_i$, $i = 1, ..., n$.

Principe du test

Le principe de calcul est très similaire au test de Kolmogorov-Smirnov, à la différence que les paramètres de la loi sont estimés. L'exécution du test de Lilliefors est donnée par les étapes suivantes :

1. ordonner les observations de manière croissante ;
2. calculer les paramètres $\overline{x}$ et s^2;
3. calculer alors les données centrées et réduites z_i;
4. obtenir les valeurs $\Phi\left(z_{(i)}\right)$;
5. calculer la valeur de la statistique L;

Pour décider, on compare la valeur L avec une certaine valeur critique L_{crit} correspondant à un seuil de signification α fixé. Si $L > L_{crit}$, l'hypothèse H_0 est rejetée avec un risque α sinon elle est acceptée. Les valeurs critiques L_{crit} sont tabulées (voir, par exemple, [6]).

2.1.5 Test de Shapiro-WilK

Le test de Shapiro-WilK est un très populaire en comparaison des autres tests. Il est particuliérement puissant pour les petits effectifs ($n \leq 50$). Il est basé sur la statistique

$$W := \frac{\left[\sum_{i=1}^{[n/2]} a_i \left(x_{(n-i+1)} - x_{(i)}\right)\right]^2}{\sum_i (x_i - \overline{x})},$$

où [.] désigne la partie entière et les a_i sont des constantes, tabulées, 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.

On rejette l'hypothèse de normalité lorsque $W > W_{crit}$. Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans le tableaux Shapiro-Wilk (voir, par exemple [15]).

2.2 Evaluation de la normalité multivariée

De nombreux tests et procédures graphiques sont proposés pour évaluer si des données multidimensionnelles provenaient d'une population normale multivariée. Il est donc impératif que des procédures existent pour détecter les cas où les données présentent des écarts modérés à extrêmes par rapport à ce qui est attendu sous normalité multivariée. Une première possibilité est de vérifier chaque variable séparément pour la normalité univariée. En d'autres termes, on veut répondre à cette question : les composantes X_j vérifient-elles ou non l'hypothèse de normalité ?

Ensuite et comme toute combinaison linéaire de v.a normales est normale et que les contours de la densité normale multivariée sont des ellipsoïdes, alors on aborde ces questions :

1- Quelques combinaisons linéaires des composants X_j sont-elles normales ?

2- Le diagramme de dispersion de paires (ou plus) d'observations sur différentes caracté-

ristiques présente-il une apparence elliptique ?

2.2.1 Evaluation de la normalité des distributions marginales

Si l'histogramme d'une variable X_j semble assez symétrique, on peut vérifier davantage en comptant le nombre d'observations à certains intervalles. Une distribution normale univariée attribue une probabilité 0.683 à l'intervalle $(\mu_j - \sigma_j, \mu_j + \sigma_j)$ et la probabilité 0.954 à l'intervalle $(\mu_j - 2\sigma_j, \mu_j + 2\sigma_j)$. Par conséquent, avec une grande taille de l'échantillon n, on s'attend à ce que la proportion $\hat{p}_{j1}$ des observations situées dans l'intervalle $(\bar{x}_j - s_j; \bar{x}_j + s_j)$ soit voisine de 0.683. De même, la proportion observée $\hat{p}_{j2}$ des observations dans $(\bar{x}_j - 2s_j, \bar{x}_j + 2s_j)$ doit être à peu près égale à 0.954. D'après [10], page 178, l'utilisation de l'approximation normale de la distribution d'échantillonnage de $\hat{p}_j$ donne

$$\mid \hat{p}_{j1} - 0.683 \mid > 3\sqrt{\frac{(0.683)(0.317)}{n}} = \frac{1.396}{\sqrt{n}},$$

et

$$\mid \hat{p}_{j2} - 0.954 \mid > 3\sqrt{\frac{(0.954)(0.046)}{n}} = \frac{0.628}{\sqrt{n}}.$$

Lorsque les proportions observées sont trop faibles, des distributions des queues plus épaisses que la normale sont suggérées.

Les graphiques sont toujours des dispositifs utiles dans toute analyse de données. Des tracés spéciaux appelés Q-Q plot sont souvent utilisés pour évaluer l'hypothèse de normalité des marginales. C'est un nuage depoints formé par les quantiles empiriques par rapport aux quantiles théoriques espérés si les observations étaient effectivement distribuées normalement. Lorsque les points de ce nuage sont approximativement alignés, l'hypothèse de normalité reste valable. La normalité est suspecte si les points s'écartent d'une ligne droite. De plus, la configuration des écarts peut fournir des indices sur la nature de la non-normalité.

Pour simplifier la notation, si $x_1, ...x_n$ représentent n observations X_j alors les observations

ordonnées $x_{(1)} \leq ... \leq x_{(n)}$ représentent les quantiles de l'échantillon. Lorsque les les $x_{(j)}$ sont distincts, exactement j observations sont inférieures ou égales à $x_{(j)}$. Pour une distribution normale standard, les quantiles théoriques $q_{(j)}$ sont définis par la relation

$$p_{(j)} := P[Z \leq q_{(j)}] = \int_{-\infty}^{q_{(j)}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{j - 1/2}{n}.$$

Si les données proviennent d'une population normale, les paires de quantiles $(q_{(j)}, x_{(j)})$ seront approximativement liées linéairement, puisque $\sigma q_{(j)} + \mu$ est (presque) le quantile espéré.

Etapes menant au Q-Q plot

1. ordonner les observations originales pour obtenir $x_{(1)} \leq ... \leq x_{(n)}$ et leurs valeurs de probabilité correspondantes $(1 - 1/2)/n, ..., (n - 1/2)/n$.

2. calculer les quantiles normaux standard $q_{(1)}, ...q_{(n)}$.

3.tracer les paires d'observations $(q_{(1)}, x_{(1)}), ..., (q_{(n)}, x_{(n)})$ et examiner la linéarité du nuage de points.

Remarque 2.2.1 *Le Q-Q plot n'est informatif que si la taille de l'échantillon est modérée à grande, par exemple* $n > 25$.

Test du coefficient de corrélation pour la normalité ([10], page 181)

Le coefficient de corrélation est défini par

$$r_Q := \frac{\sum_{j=1}^{n} (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n} (x_{(j)} - \bar{x})^2} \sqrt{\sum_{j=1}^{n} (q_{(j)} - \bar{q})^2}}. \tag{2.5}$$

La région critique : on rejette l'hypothèse de normalité lorsque

$$r_Q \leq r_{crit},$$

où r_{crit} est une valeur critique tabulée, pour différents risques α et effectifs n, dans le tableau 2.2.

Taille de l'échantillon	Niveau de signification α		
n	0.01	0.05	0.10
5	0.8299	0.8788	0.9032
10	0.8801	0.9198	0.9351
15	0.9126	0.9389	0.9503
20	0.9269	0.9508	0.9604
100	0.9822	0.9873	0.9895
150	0.9879	0.9913	0.9928
200	0.9905	0.9931	0.9942

TAB. 2.2 – Valeurs critiques du test du coefficient de corrélation

2.2.2 Évaluation de la normalité conjointe

Les distributions marginales univariées étaient considérées plus haut, on s'intéresse au cas multivarié. On a décrit les diagrammes de dispersion pour des paires de caractéristiques. Si les observations étaient générées à partir d'une distribution normale multivariée, chaque distribution univariée serait normale et les contours de densité constante seraient des ellipses. Le nuage de points doit se conformer à cette structure en présentant un modèle global presque elliptique.

Le résultat (1.3) permet de dire que l'ensemble des résultats tels que

$$(X-\mu)^{t}\Sigma^{-1}(X-\mu) \leq \mathcal{X}_p^2(\alpha),$$

a une probabilité égale à $1-\alpha$. De plus, d'après (1.4), on doit s'attendre (de façon empirique) à peu près au même pourcentage de $(1-\alpha)\times 100\%$ d'observations se trouvant dans l'ellipse donnée par

$$(X-\bar{X})^{t}S^{-1}(X-\bar{X}) \leq \mathcal{X}_p^2(\alpha). \tag{2.6}$$

Sinon, il y a un doute sur la normalité du V.a X. Les valeurs seuils $\mathcal{X}_p^2(\alpha)$ sont lues dans la table du Khi-deux pour différents risques α et effectifs n (voir, par exemple [12]).

Test du Khi-deux

C'est une méthode plus formelle pour juger de la normalité conjointe d'un ensemble de données vectorielles. Elle est basée sur les distances généralisées

$$d_{(i)}^2 := (X^{(i)} - \bar{X})^t S^{-1} (X^{(i)} - \bar{X}),\ i = 1, ..., n. \tag{2.7}$$

En pratique, d'après (1.4), lorsque la population parente est normale et que n et $(n-p)$ sont tous les deux supérieurs à 25, chacune des $d_{(i)}^2$ doit se comporter comme une v.a du Khi-deux. Le graphique résultant est appelé graphe du Khi-Deux.

La région critique : on rejette l'hypothèse de normalité lorsque

$$d_{(i)}^2 > \mathcal{X}_p^2(\alpha).$$

Etapes de construction du graphe du Khi-deux

1. ordonner les $d_{(i)}^2$ du plus petit au plus grand comme $d_{(1)}^2 \leq .... \leq d_{(n)}^2$;

2. représenter graphiquement les paires $(q_p(\frac{i-1/2}{n}), d_{(i)}^2)$, où $q_p(\frac{i-1/2}{n}) = \mathcal{X}_p^2(\,(n-i+1/2)\,/n)$ est le quantile $(i-1/2)\,/n$ de la distribution du Khi-deux avec p ddl.

Le tracé doit avoir une allure droite le long de la première bissectrice. Un nuage courbé suggère un manque de normalité.

2.3 Application sous R

Dans cette partie, on applique la méthode graphique du Q-Q plot et quelques tests de normalité aussi bien dans le cas univarié que multivarié. On commence par simulation de la loi normale multivariée qui nous permet de présenter ces propriétés. Ensuite, on

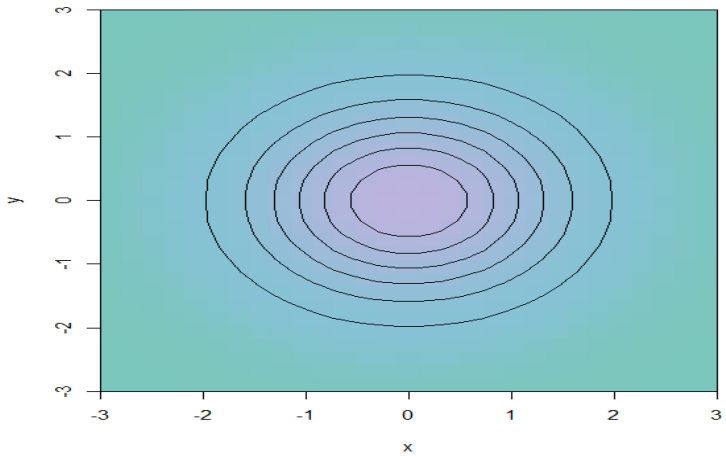

FIG. 2.1 – Densité de la loi normale bivariée en 2D.

applique les tests de normalité qu'on a présenté plus haut. Les résultats numériques et les représentations graphiques sont obtenus au moyen du logiciel de traitement et analyse statistiques **R**, introduit par R. Ihaka et R. Gentlman [7]. Pour une brève description de ce logiciel, voir l'annexe A.

2.3.1 Simulation de la loi normale bivariée

On présente les graphes de la densité et de la fonction de répartition de la loi normale bivariée ($p = 2$) pour différentes valeurs de n.

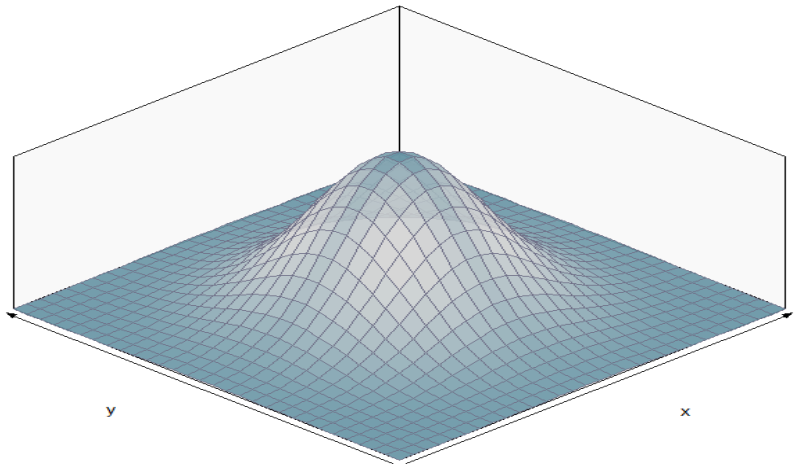

FIG. 2.2 – Densité de la loi normale bivariée en 3D.

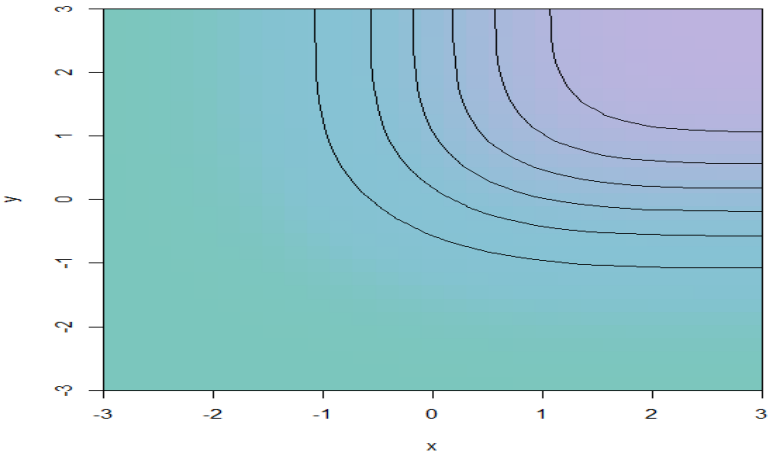

FIG. 2.3 – Fonction de répartition de la loi normale bivariée en 2D.

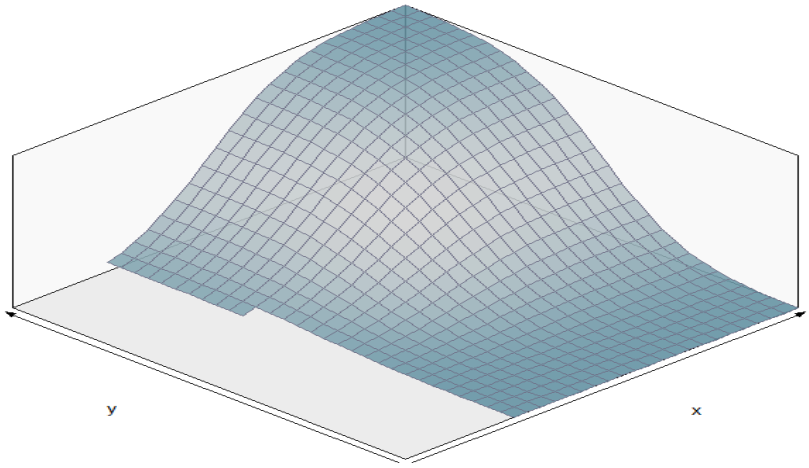

FIG. 2.4 – Fonction de répartition de la loi normale bivariée en 3D.

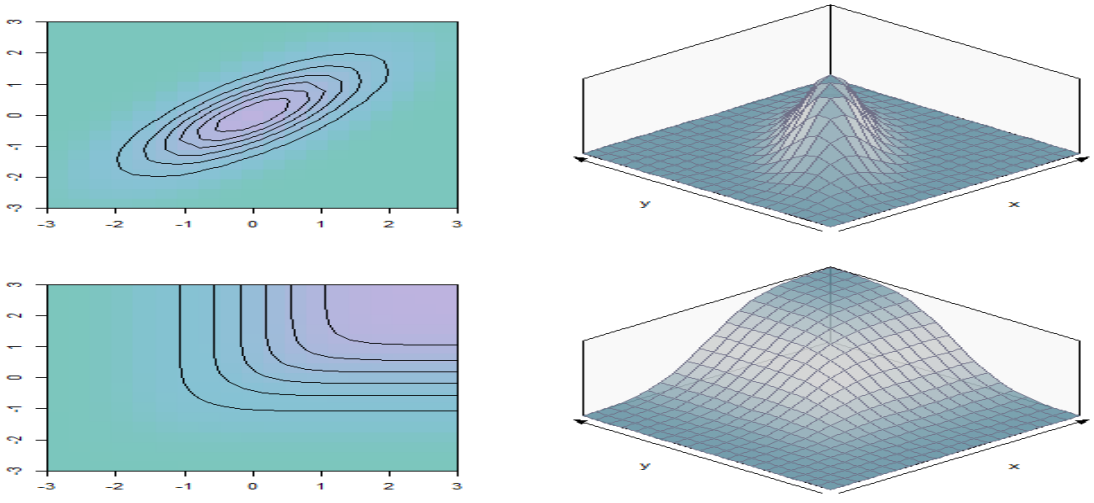

FIG. 2.5 – Densité et fonction de répartition pour $n = 20$.

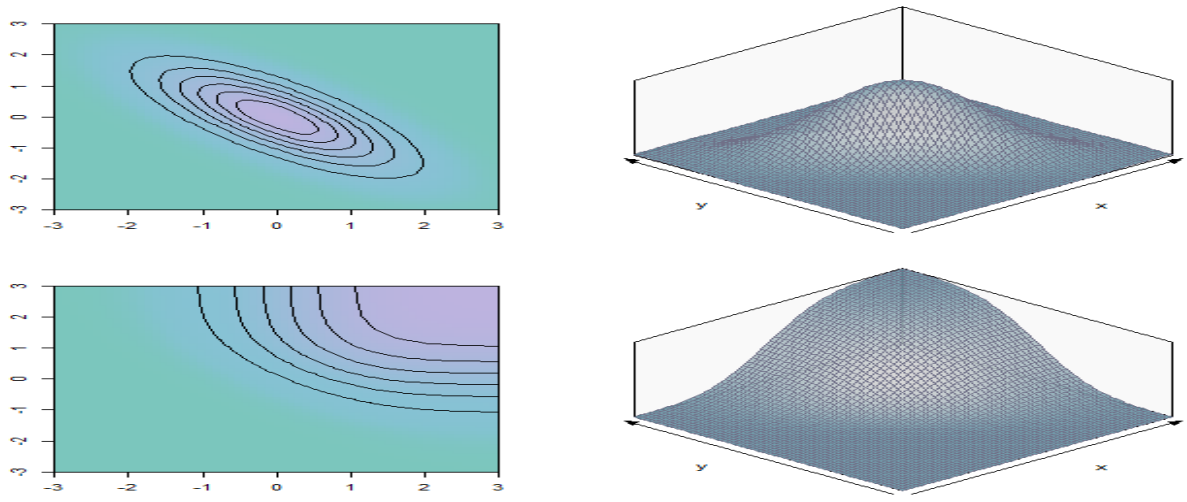

FIG. 2.6 – Densité et de fonction de répartition pour $n = 60$.

2.3.2 Tests de normalité univariée

Exemple 2.3.1 (tests usuels)

Choix du test : *Kolomogrov-Smirnov, Cramer-von Mises, Shapiro-Wilk, Anderson-Darling et Lilliefors.*

Choix des lois : *on simule deux lois différentes de la loi normale : la loi exponentielle de paramètre* $\lambda = 2$ *et celle de Student de paramètre* 30.

Choix de la taille : *on choisit différentes tailles pour conclure s'il existe un effet de taille sur les résultats du test. On prend* $n = 10$ *et* $n = 50$. *Dans les tableaux suivants, on compare les p-values de tous les tests avec* $\alpha = 0.01$, *c-à-d*

$$\begin{cases} \text{si } p\text{-value} > 0.01, \ \text{en accepte} H_0, \\ \text{si } p\text{-value} \leq 0.01, \ \text{en rejette } H_0. \end{cases}$$

	cvm	$p-v$	sw	$p-v$	ad	$p-v$	ks	$p-v$	lil	$p-v$
$\mathcal{N}$	0.059	0.346	0.966	0.859	0.482	0.177	0.000	1.000	0.226	0.154
$\mathcal{E}$	0.060	0.340	0.760	0.004	0.377	0.330	0.400	0.410	2.92	0.020
t	0.020	0.040	0.903	0.540	0.237	0.840	0.500	0.160	0.124	0.920

TAB. 2.3 – Résultats des tests de normalité pour n=10

Commentaire : *d'après le tableau 2.3, on observe que :*

1. *Le test de Shapiro-Wilk, de p-value très petite inférieure à* 0.01, *est le seul test qui rejette l'hypothése de normalité pour la loi exponentielle.*

2. *Tous les tests donnent de bons résultats pour la loi de Student.*

	cvm	$p-v$	sw	$p-v$	ad	$p-v$	ks	$p-v$	lil	$p-v$
$\mathcal{N}$	0.070	0.200	0.970	0.530	0.338	0.480	0.000	1.000	0.080	0.490
$\mathcal{E}$	0.500	0.000	0.840	0.000	1.940	0.000	0.400	0.001	0.160	0.001
t	0.060	0.288	0.960	0.170	0.773	0.040	0.180	0.390	0.810	0.510

TAB. 2.4 – Résultats des tests de normalité pour n=50

Commentaire : *d'après le tableau 2.4, on observe que :*

1. *Tous les tests rejettent la normalité pour la loi exponentielle.*

2. *Tous les tests donnent des bons résultats pour la loi de Student.*

Exemple 2.3.2 (Q-Q plot [10], page 179) *Un échantillon de* $n = 10$ *de valeurs simulées sont ordonnées dans le tableau 2.5.*

observations ordonnées $x_{(j)}$	niveaux de probabilité $(j-1/2)/n$	standard normal quantiles $q_{(j)}$
-1	0.05	-1.645
-0.1	0.15	-1.036
0.16	0.25	-0.674
0.41	0.35	-0.385
0.62	0.45	-0.125
0.80	0.55	0.125
1.26	0.65	0.385
1.54	0.75	0.674
1.71	0.85	1.036
2.30	0.95	1.645

TAB. 2.5 – Tableau de données.

Le Q-Q plot pour les données précédentes est illustré à la figure 2.7.

Commentaire : *d'après la figure 2.7 on remarque que les paires de points* $\left(q_{(j)}, x_{(j)}\right)$ *se trouvent presque le long d'une ligne droite. Donc, on ne rejette pas l'idée que ces données sont normalement distribuées - en particulier avec une taille d'échantillon aussi petite que* $n = 10$.

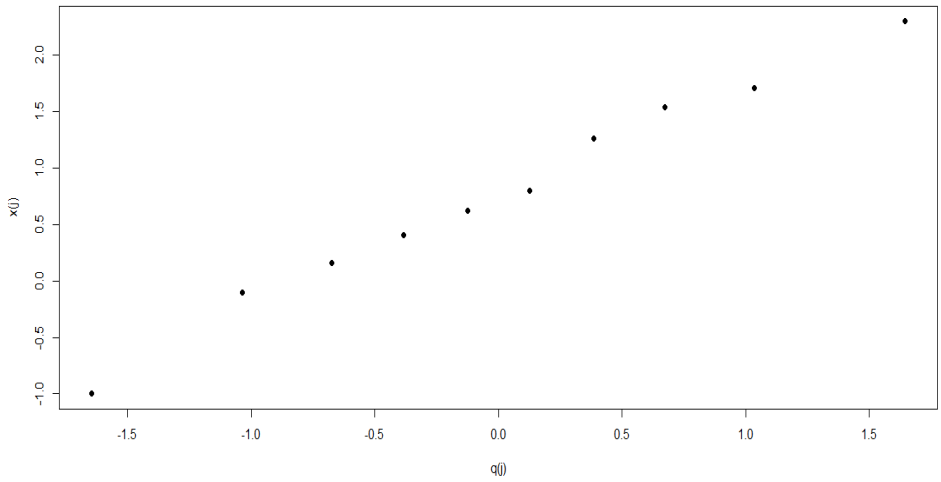

FIG. 2.7 – Q-Q plot pour des données de taille $n = 10$.

Exemple 2.3.3 (test du coefficient de corrélation [10], page 182) *En utilisant les informations du tableau 2.5 de l'exemple précédent, on a* $\bar{x} = \frac{1}{n}\sum_{j=1}^{10} x_{(j)} = 0.77$ *et* $\sum_{j=1}^{10}(x_{(j)} - \bar{x})q_{(j)} = 8.584$, *et* $\sum_{j=1}^{10}(x_{(j)} - \bar{x})^2 = 8.472$, *et* $\sum_{j=1}^{10} q_{(j)}^2 = 8.795$.

D'après (2.5), on calcule le coefficient de corrélation

$$r_Q = \frac{8.584}{\sqrt{8.472}\sqrt{8.795}} = 0.994.$$

Commentaire : *pour* $n = 10$ *et* $\alpha = 0.10$, *on a* $r_{crit} = 0.935$. *Puisque* $r_Q > r_{crit}$, *alors on ne peut pas rejter l'hypothèse de normalité au niveau de signification* 10%. *Ceci confirme la conclusion ci-dessus faite au vu du Q-Q plot de la figure 2.7.*

Exemple 2.3.4 (Q-Q plot [10], page 180) *Des observations du rayonnement émis à travers les portes fermées de* $n = 42$ *fours choisis au hasard ont été faites. Ces données réelles sont répertoriées dans le tableau 2.6.*

A partir des paires $(q_{(j)}, x_{(j)})$, *on construit le Q-Q plot, illustré par la figure 2.8.*

Commentaire : *il ressort du graphique que les données dans leur ensemble ne sont pas normalement distribuées. Les points finaux du nuage sont des valeurs aberrantes trop*

$x_{(j)}$	$q_{(j)}$	$x_{(j)}$	$q_{(j)}$	$x_{(j)}$	$q_{(j)}$
0.01	−2.326	0.09	−1.282		
0.01	−2.326	0.09	−1.282	0.20	−0.842
0.02	−2.054	0.10	−1.282	0.20	−0.842
0.02	−2.054	0.10	−1.282	0.30	−0.524
0.02	−2.054	0.10	−1.282	0.30	−0.524
0.03	−1.881	0.10	−1.282	0.30	−0.524
0.05	−1.645	0.10	−1.282	0.30	−0.524
0.05	−1.645	0.10	−1.282	0.40	−0.253
0.05	−1.645	0.10	−1.282	0.40	−0.253
0.05	−1.645	0.10	−1.282		
0.05	−1.645	0.10	−1.282		
0.07	−1.476	0.11	−1.227		
0.08	−1.405	0.12	−1.175		
0.08	−1.405	0.15	−1.036		
0.08	−1.405	0.15	−0.915		
0.09	−1.341	0.18	−0.915		
0.09	−1.341	0.20	−0.842		

TAB. 2.6 – Observations ordonnées et quantiles correspondants

grandes par rapport au reste des observations.

2.3.3 Test de normalité bivariée

Exemple 2.3.5 ([10], page 183) *Les données réelles du tableau 2.7 constituent des paires d'observations x_1 (vente) et x_2 (bénéfice) pour les 10 plus grandes compagnies du monde. On applique le test de normalité bivariée du Khi-deux.*

On a

$$\overline{x} = \begin{pmatrix} 155.06 \\ 14.70 \end{pmatrix}, \; S = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}, \; S^{-1} = \begin{bmatrix} 0.00025 & -0.00293 \\ -0.00293 & 0.07215 \end{bmatrix}.$$

Sur la table du khi-deux on lit $\chi^2_2(0.5) = 1,39$. *Toute observation* $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}\right)^t$, $i = 1, ..., 10$, *satisfaisant la condition* (2.6), *c-à-d*

$$\begin{pmatrix} x_1^{(i)} - 155.06 \\ x_2^{(i)} - 14.70 \end{pmatrix}^t \begin{pmatrix} 0.000253 & -0.002930 \\ -0.002930 & 0.072148 \end{pmatrix} \begin{bmatrix} x_1^{(i)} - 155.06 \\ x_2^{(i)} - 14.70 \end{bmatrix} \leq 1.39$$

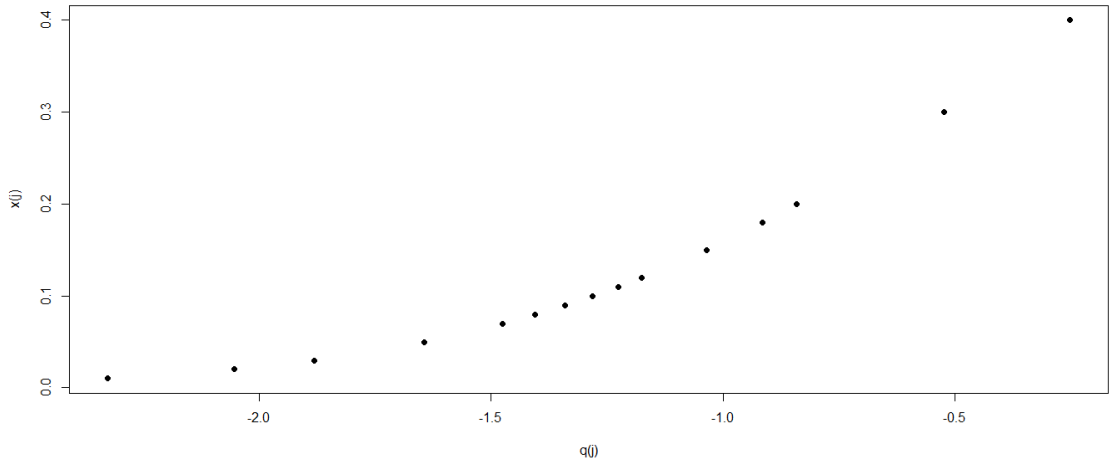

FIG. 2.8 – Q-Q plot des données de rayonnement de fours

compagnie	x_1 (milliards)	x_2 (milliards)
Citigroup	108.28	17.05
General Electric	152.36	16.59
American Inti Group	95.04	10.91
Bank of America	65.45	14.14
HSBC Group	62.67	9.52
Exxon Mobil	263.99	25.33
Royal Dutch/Shell	265.19	18.54
BP	285.06	15.73
lNG Group	92.01	8.10
Toyota Motor	165.68	11.13

TAB. 2.7 – Les 10 plus grandes entreprises du monde

est sur ou à l'intérieur de l'ellipse à 50%. *Sinon, l'observation est en dehors de ce contour. Les distances généralisées par rapport à* $\overline{x}$, *calculées à partir de* (2.7) , *pour les* 10 *paires d'observations sont résumées dans le tableau* 2.8.

Commentaire : *on constate que parmi les* 10 *paires,* 4 *ont des distances généralisées inférieures au seuil* 1.39. *Ce qui veut dire qu'une proportion de* 40% *des données se situe dans le contour de* 50%. *Si les observations étaient normalement distribuées, alors environ la moitié d'entre elles se trouveraient dans ce contour. Cette différence de proportions peut normalement nous amener à rejeter la normalité bivariée.*

$d^2_{(1)}$	$d^2_{(2)}$	$d^2_{(3)}$	$d^2_{(4)}$	$d^2_{(5)}$	$d^2_{(6)}$	$d^2_{(7)}$	$d^2_{(8)}$	$d^2_{(9)}$	$d^2_{(10)}$
1.60	0.30	0.62	1.79	1.30	4.38	1.64	3.53	1.71	1.16

TAB. 2.8 – Distances généralisées pour les données des compagnies

Construction du graphique du Khi-deux

Les distances ordonnées et les quantiles du Khi-deux correspondants à $p = 2$ et $n = 10$ sont donnés dans le tableau 2.9. Le graphe du Khi-deux est illustré par figure 2.9.

$d^2_{(i)}$ ordonnée	0.30	0.62	1.16	1.30	1.61	1.64	1.71	1.79	3.53	4.38
$q_p(\frac{i-1/2}{n})$	0.10	0.33	0.58	0.86	1.20	1.60	2.10	2.77	3.79	5.99

TAB. 2.9 – Distances généralisées et quantiles du Khi-deux

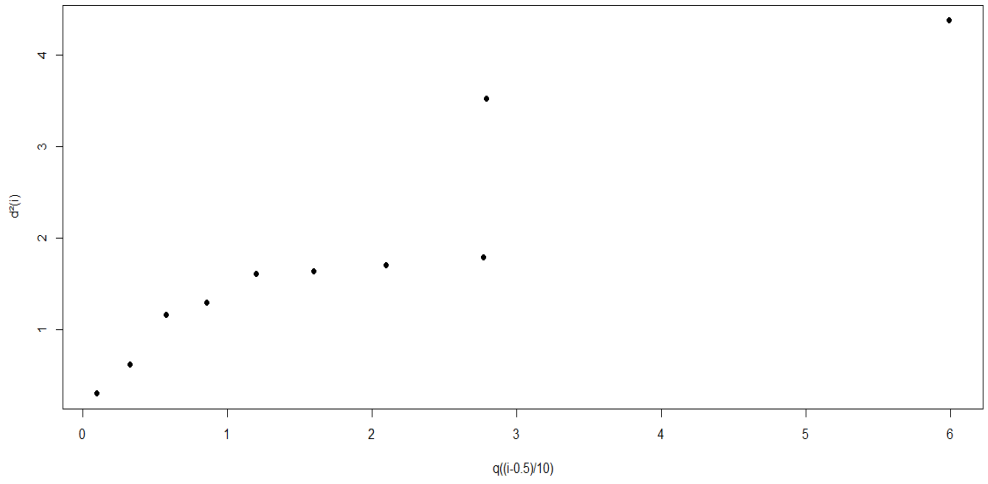

FIG. 2.9 – Q-Q plot du Khi-deux pour les distances ordonnées.

Commentaire : *le nuage de points de la figure 2.9 semble assez aligné. Compte tenu de la petite taille de l'échantillon, il est difficile de rejeter la normalité bivariée sur la base de ce graphique.*

Conclusion

Le but de ce mémoire est de présenter les défférentes méthodes des tests de normalité multivariée. Il existe deux méthode pour vérifier la normalité. La première méthode est la méthode graphique où elle nous donne une ideé sur la distribution de l'échantion réel telle que : Le graphique Q-Q plot pour les données, qui est un graphique des données ordonnées $x_{(j)}$ par rapport aux quantiles normaux $q_{(j)}$, se trouvent presque le long d'une ligne droite, et nous ne rejetterions pas l'idée que ces données sont normalement distribué, et le graphique du Khi-deux des distances ordonnées. Cependant, la méthode graphique n'est pas suffisante pour conclure de la normalité, pour cette raison on réalise les tests d'hypothèses. Parmi ces derniers, on peut citer le test du coefficient de corrélation pour la normalité et celui basé sur les distances d_j^2. On résume les étapes des tests de normalité dans les points suivants :

1. on choisit un test parmi les tests de normalité ;
2. on fixe une valeur pour le niveau de signification α;
3. on calcule la statistique correspendant au test choisi ;
4. on compare le résultat obtenu le point critique du test au seuil α;
5. on prend la décision : rejet ou non de la normalité.

Bibliographie

[1] A -lefrance., (2006). Compléments de cours de probabilités. DEUG $2^{ème}$ annés.

[2] D'Agostino, R. B. & Stephens, M. A. (1986). Goodness-of-fit techniques. Marcel Dekker, Inc.

[3] Mohamed, C & Ali, G. et jérome, S. (2008). Estimation de quantiles, géométrique conditionnels et non conditionnels. Université de Bordeaux.

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

[5] Ginsbourger, D & Bay,X. (2009).Cours de Processus Al´eatoires. Département 3MI, Ecole des Mines de Saint-Etienne

[6] http ://courses.wcupa.edu/rbove/eco252/252KStest.doc.

[7] Ihaka, R., Gentleman, R. (1996). R : A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5 : 299 − 314.

[8] Mouchiroud, D. (2002). Variable aléatoire. Oukomputile pour la Biologie- Deug SVI-UCBI.

[9] Jean,M (2006). Probabilité et statistique. Départemk-e genie industriale. INSA Lyon1.

[10] Johnson, R. & Wichern, D. (2007). Applied multivariate statistical analysis_Pearson Prentice Hall.

[11] Rencher, A. (2002). Methods of Multivariate Analysis_Wiley.

[12] Saporta, G. (2006). Probabilitè, analyse de données et statistique. Technip.

[13] Singla, N. Jain, K. & Sharma, S. K. (2016). Goodness of fit tests and power comparisons for weighted gamma distribution. REVSTAT-Statistical Journal, 14(1), $29-48$.

[14] Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American statistical Association, 69 (347), $730-737$.

[15] Rakotomalala, R. Tests de normalité Techniques empiriques et tests statistiques. Université Lumière Lyon 2.

Annexe A : Quelques éléments de R

Le logiciel **R** est un language de programmation et un environnement mathématique utilisé pour le traitement de données et l'analyse statistique. Il existe plusieurs versions, telle que pour l'application de ce mémoire on utilisé la version **Rx64 4.0.2.** Ce logiciel contient des packages de base trouvés dans toute les versions, et des packages correspond avec quelque versions, telle que le package **(nortest)** qui contient les test de normalité par exemple Cramer- von Mises, Shapiro-Wilk, Anderson-Darling, et Lillieforse. et le package **(mvtnorm)** qui contient des fonctions permettant de manipuler facilement les vecteurs gaussiens, le package **(bivariate)** pour évaluer les distributions normales bivariées.

Fonction nous permettent de générer un échantillion

rexp(x) : pour générer un échantillion de loi exponentille.

rpoios(x) : pour générer un échantillion de loi poisson.

rt(x) : pour générer un échantillion de loi student.

rnorm(x) : pour générer un échantillion de loi normale.

rchisq(x) : pour générer un échantillion de loi Khi-deux.

Exemples

x=rnorm$(50, 1, 2)$

x1=rpois$(50, 2)$

x2=rexp$(50, 2)$

x3=rt$(50, 2)$

x4=rchisq$(50, 2)$

Fonction nous permettent de calcul le quantile un échantillion

qnorm : Fonction nous permettent de calcul le quantile de la loi normale.

qrchisq : Fonction nous permettent de calcul le quantile de la loi Khi-deux.

Fonction nous permettent de calcul la densité bivariée

dmvnorm : permettent de calcul la densité bivarié

nbvpdf : fonction de probabilité de la loi normale bivariée.

nbvcdf : fonction comulative la loi normale bivariée.

Exemples

dmvnorm : (c(0, 0), c(0, 0), diag(2), log=FAlSE)

$f1 < -$nbvpdf(0, 0, 1, 1, 0),

$F < -$nbvcdf(0, 0, 1, 1, 0),

Fonction nous permettent de définir les méthodes graphique

plot(x,y) : La commande plot permet de tracer un ensemble de points de coordonnées (x_i, y_i).

qqplot(q,x) : Graph de qantiles de x avec les valeurs attendues selon la loi normale.

Fonctions nous permettent de faire les test de normalité

Ks.test : Test de kolmogrov-smirnov.

cvm.test : Test de cramer von-mises.

shapiro.test : Test de shapiro wilk.

ad.test : Test d'anderson-Darlling.

lillie.test : Test de lilliefors.

cor.test : Test de coefficient corrélation.

Exemples

>x=rnorm(500,1,2)

>ks.test(x,x)

Two-sample Kolmogorov-Smirnov test

data : x and x

D = 0, p-value = 1

>cvm.test(x)

Cramer-von Mises normality test

data : x

> ad.test(x)

W = 0.0537, p-value = 0.457

Anderson-Darling normality test

data : x

A = 0.2959, p-value = 0.5932

> shapiro.test(x)

Shapiro-Wilk normality test

data : x

W = 0.9984, p-value =0.9232

> lillie.test(x)

Lilliefors (Kolmogorov-Smirnov) normality test

data : x

L = 0.0177, p-value = 0.9636

>cor.test(x,q)

Pearson's product-moment correlation

data : x and q

$t = 17.24, df = 39, p - value < 2.2e - 16$

alternative hypothesis : true correlation is not equal to 95 percent confidence interval :

$r_Q = 0.8899$, p-value=0.9678

Cov(A) : Fonctions permet de caclculer la matrice de covaraince .

Solve(A) : Fonctions permet de caclculer l'inverse de la matrice A.

det(A) : Fonctions permet de caclculer le déterminant de la matrice A .

Annexe B : Abréviations et Notations

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

A_n^2	statistique Anderson-Darling.
A_{crit}^2	saleur critique d'Anderson-Darling.
α	risque de premier espèce.
c-à-d	c'est à dire.
ddl	degré de liberté.
D_n	statistique de Kolomogrov-Smirnov.
D_{crit}	valeur critique de Kolomogrov-Smirnov.
$E[X]$	espérance mathématique ou moyenne d'une v.a X.
$\mathcal{E}$	loi de exeponentielle.
F	fonction de répartition.
F_n	fonction de répartition empirique.
f	densité de probabilité.
Φ	fonction de répartition de $\mathcal{N}(0;1)$.
ϕ	fonction caractéristique.
iid	indépendantes identiquement distribuées.

H_0	hypothèse nulle.
H_1	hypothèse alternative.
L	statistique de Lilliefors.
L_{crit}	valeur critique de Lilliefors.
$\max(A)$ ou $(\min(A))$	maximum de A (ou minimum de A)
$\mathcal{N}(0,1)$	loi normale centrée réduite.
q	quantile.
r_Q	coefficient de corrélation.
r_{crit}	valeur critique de r_Q.
S_n^2	variance empirique.
t	loi de Student.
$Var(X)$	variance de X.
v.a	variable aléatoire.
V.a	vecteur aléatoire.
v.a.c	variable aléatoire continue.
$\overline{w}_n^2$	statistique de Cramer-von Mises.
$\overline{w}_{crit}^2$	valeur critique de Cramer-von Mises.
$\overline{X}$	moyenne empirique.
$(X_1, ..., X_n)$	échantillon de taille n de X.
$\mathcal{X}_p^2$	loi de khi-deux.
Σ	matrice de covaraince.
$\amalg_A$	fonction indicatrice de l'ensemble A.
W	statistique de Shapiro-Wilk.
W_{crit}	valeur critique de Shapiro-Wilk.
$\sim$	suit la loi.
$\rightsquigarrow$	suit approximativement la loi.
$:=$	égalité par définition.

ملخص

الهدف من هذه المذكرة هو معرفة اختبارات التوزيع الطبيعي الذي نستطيع بواسطته معرفة قانون عينة متعددة المتغيرات إذا كانت تابعة لقانون التوزيع الطبيعي متعدد المتغيرات أم لا . في هذه الأطروحة نقدم القانون العادي متعدد المتغيرات وخصائصه والطرق التي تسمح لنا بفحص الحالة الطبيعية، الطريقة الأولى هي الطريقة البيانية وتستخدم البيان (مؤامرة الكميات) ولكن هذه الطريقة تعطينا فكرة حول خصائص التوزيع الطبيعي للمتغيرات. الطريقة الثانية هيا اختبارات التوزيع الطبيعي في الحالة الفردية والحالة متعددة المتغيرات حيث قمنا بإعطاء كل اختبار الإحصائية الخاصة به مع منطقة الرفض والقبول.

الكلمات المفتاحية: قانون التوزيع الطبيعي، مؤامرة الكميات، مسافات مربعة معممة، معامل الارتباط، قيمة.

Abstract

Multivariate normality tests are goodness-of-fit tests, which verify whether a set of multivariate observations could come from a multivariate normal distribution or not. In this dissertation, we present the multivariate normal distribution and its properties and the methods which allow us to check the normality. The first method is a graphical one (Q-Q plot), which gives an idea about the shape of the distribution. The second method is the normality tests in the univariate and multivariate case (Kolmogorov-Smirnov, Cramer-Von Mises, Anderson Darling, correlation coefficients, squared generalized distances...), where we gave the statistic of each test and the critical region.

***Keywords**: Normality tests, Q-Q plot, squared generalized distances, correlation coefficient, p-value.*

Résumé

Les tests de normalité multivariées sont des tests d'ajustement, qui permettent de vérifier si un ensemble d'observations multivariées pourrait provenir d'une distributions normale multivariée ou non. Dans ce mémoire, on a présenté la loi normale multivariée et ses propriétés et les méthodes qui nous permettent de vérifier la normalité. La première méthode est une méthode graphique (Q-Q plot) qui nous donne une idée sur la forme de la distribution. La deuxième méthode est les tests de normalité dans le cas univariée et multivariée (Kolmogorov-Smirnov, Cramer-Von Mises, Anderson Darling , distances généralisées au carré,...), où on a donné la statistique de chaque test et la région critique .

***Mots clés:** Tests de normalité, Q-Q plot, distances généralisées au carré, coefficient de corrélation, p-value.*