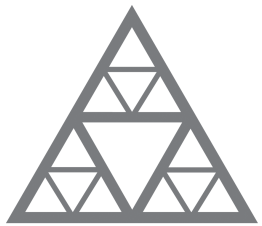
Shapiro Wilk Test and QQ-Plot

Statistics and Data Analysis

Ecoles des Ponts ParisTech

January, 2022



École des Ponts
ParisTech

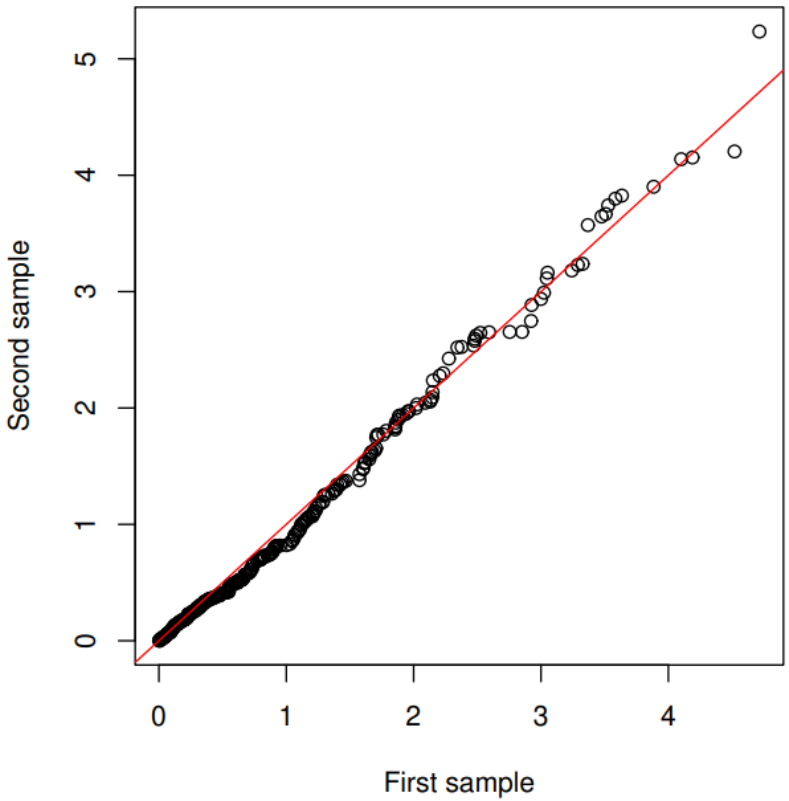
Soit X et Y deux variables aléatoires à valeur réelle et F et G leur fonction de répartition respective. Le diagramme Quantile-Quantile (Q-Q) de F et G est définie comme la courbe paramétrique

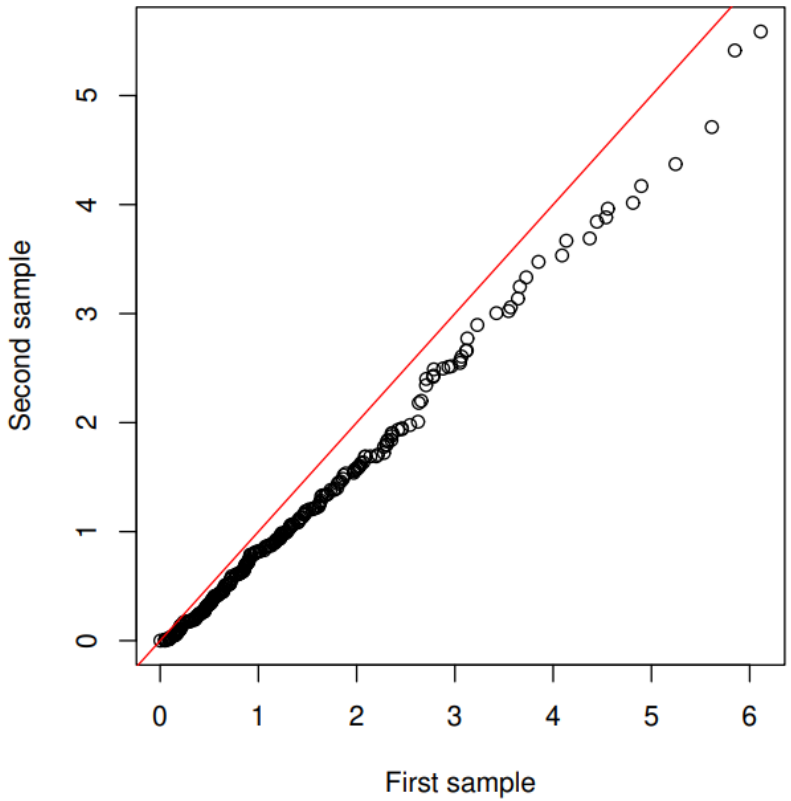
$$u \in (0,1) \rightarrow \left(F^{-1}(u), G^{-1}(u)\right)$$

F^{-1} et G^{-1} étant les pseudo-inverse de F et G.

Le diagramme Q-Q est donc supporté par la diagonale $x = y$ si et seulement si $F = G$.
Cette équivalence peut s'avérer utile dans deux contextes:

- Vérifier la qualité d'ajustement d'un échantillon $X_1, ..., X_n$ avec hypothèse nulle $F = F_0$
- Pour un test d'homogénéité de deux échantillons


Second sample
First sample
0
1
2
3
4
5


Second sample
First sample
0
1
2
3
4
5
6

Lemme

La variable aléatoire est gaussienne si et seulement si le diagramme Q-Q de sa fonction de répartition avec celle d'une loi normale centrée réduite est une ligne droite. Plus précisément, $Y \sim \mathcal{N}(\mu, \sigma^2)$ *si est seulement si le diagramme Q-Q a pour équation* $y = \sigma x + \mu$.

Démonstration du lemme

- $X \sim \mathcal{N}(0,1)$
- Supposons que $Y \sim \mathcal{N}(\mu, \sigma^2)$
- $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\sigma X + \mu \leq \sigma x + \mu) = \mathbb{P}(Y \leq \sigma x + \mu) = G(\sigma x + \mu)$
- Donc pour $u = F(x) \in [0\ 1]$, $G^{-1}(u) = G^{-1}(G(\sigma x + \mu)) = \sigma x + \mu = \sigma F^{-1}(u) + \mu$
- D'où le QQ-plot de F et G est une droite d'équation $y = \sigma x + \mu$

Démonstration du lemme

- Réciproquement, si le QQ-plot de F et G est une droite d'équation $y = \sigma x + \mu$:
- Donc pour $u = F(x) \in [0\ 1], G^{-1}(u) = \sigma F^{-1}(u) + \mu = \sigma x + \mu$
- D'où $G(\sigma x + \mu) = F(x)$ i.e. $G(y) = F(\frac{y-\mu}{\sigma})$ donc $Y \sim \mathcal{N}(\mu, \sigma^2)$

Cadre

On se place dans le cadre suivant :

- On note $Y_1, ..., Y_n$ nos observations iid
- On donne les hypothèses suivantes : $H_0 = \{P \in \mathcal{P}_0\}$ $H_1 = \{P \notin \mathcal{P}_0\}$ où $\mathcal{P}_0 = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$
- On définit ensuite $X_1, ..., X_n$ par $X_i = \frac{Y_i - \mu}{\sigma}$
- On note le vecteur $X_{()}$ comme étant la réorganisation croissante de X, $X_{(1)} \leq ...X_{(n)}$ et de même pour $Y_{()}$
- On définit ensuite $m \in \mathbb{R}^n$ par $m_i = \mathbb{E}[X_{(i)}]$ et V comme la matrice de covariance de $(X_{(1)}, ..., X_{(n)})$

Cadre

- Ceci nous permet d'écrire : $Y_{(i)} = \mu + \sigma m_i + \epsilon_i$ où $\epsilon = (\epsilon_1, ..., \epsilon_n) \sim \mathcal{N}(\mu, \sigma^2 V)$
- Puisqu'on peut approximer m_i par $\Phi^{-1}(\frac{i}{n})$, les termes $\epsilon_1, ..., \epsilon_n$ mesurent la distance entre le QQ-plot de l'échantillon $Y_{(1)}, ..., Y_{(n)}$ et la droite d'équation $y = \sigma x + \mu$
- Il reste à calculer les coefficients μ et σ ce qui correspond à un problème de régression linéaire généralisé car la matrice de covariance de ϵ n'est pas diagonale.

BLUE PROPOSITION

Proposition

Les meilleurs estimateurs linéaires de μ et de σ pour la régression linéaire de $Y_{(1)}, ..., Y_{(n)}$ sur $m_1, ..., m_n$ est

$$\hat{\mu} = \sum_{i \leq n} u_i Y_{(i)}, \quad \hat{\sigma} = \sum_{i \leq n} v_i Y_{(i)}$$

Où les vecteurs $u,v \in \mathbb{R}^n$ sont donnés par

$$u = \frac{V^{-1}1}{1V^{-1}1}, \quad v = \frac{V^{-1}m}{m^T V^{-1} m}$$

Borne pour la Statistique de Shapiro

Lemme

Presque sûrement, on a

$$\frac{\hat{\sigma}^2}{S_n^2} \leq (n-1)\|v\|^2$$

avec S_n^2 l'estimateur habituel de la variance

$$S_n^2 = \frac{1}{n-1}\sum_{1\leq n}(Y_i - \bar{Y}_n)^2$$

Démonstration du lemme

- Supposons $1^T V^{-1} m = 0$, on le montrera ensuite
- Alors, on en déduit $\sum_{i \leq n} v_i = 0$
- Donc $\hat{\sigma} = \sum_{i \leq n} v_i (Y_{(i)} - \bar{Y}_n)$ ou encore $\hat{\sigma} = \langle v, Y_{()} - \bar{Y}_n 1 \rangle$
- Ensuite par Cauchy-Schwarz, $\hat{\sigma}^2 \leq \|v\|^2 \|(Y_{()} - \bar{Y}_n)\|^2$
- D'où $\hat{\sigma}^2 \leq \|v\|^2 (n-1) S_n^2$

Démonstration de $1^T V^{-1} m = 0$ (1)

- On remarque d'abord que $(X_1, ..., X_n)$ suit la même loi que $(-X_1, ..., -X_n)$ car les X_i suivent une loi gaussienne centrée
- En réorganisant, on a $(X_{(1)}, ..., X_{(n)})$ suit la même loi que $(-X_{(n)}, ..., -X_{(1)})$
- Donc pour tout i,j $m_i = -m_{n-i+1}$ et $V_{i,j} = V_{n-i+1,n-j+1}$
- Alors $Jm = m$ et $JVJ = V$ où J$= \begin{pmatrix} 0 & ... & 0 & 1 \\ 0 & ... & 1 & 0 \\ 0 & \vdots & \vdots & \vdots \\ 1 & ... & 0 & 0 \end{pmatrix}$

Démonstration de $1^T V^{-1} m = 0$ (2)

- Or $J^{-1} = J$ donc $VJ^{-1} = VJ = J^{-1}V$ donc J et V commutent, donc J et V^{-1} aussi
- Alors $JV^{-1}m = V^{-1}Jm = -V^{-1}m$
- En réecrivant terme à terme, on a pour tout i, $(V^{-1}m)_i = -(V^{-1}m)_{n-i+1}$
- Donc $\sum_{i \leq n} (V^{-1}m)_i = 1^T V^{-1} m = 0$

Statistique de Shapiro

En repartant du lemme précédent, on a :

$$0 \leq \frac{\hat{\sigma}^2}{(n-1)\|v\|^2 S_n^2} \leq 1$$

Or

$$(n-1)S_n^2 = \sum_{i \leq n} (Y_i - \bar{Y}_n)^2$$

Et

$$\frac{\hat{\sigma}^2}{\|v\|^2} = (\sum_{i \leq n} \frac{v_i}{\|v\|} Y_{(i)})^2$$

Statistique de Shapiro

Definition

On définit la statistique de Shapiro comme

$$W = \frac{(\sum_{i \leq n} a_i Y_{(i)})^2}{\sum_{i \leq n}(Y_i - \bar{Y}_n)^2}, \qquad a = \frac{V^{-1}m}{|V^{-1}m|}$$

Donc, on obtient

$$0 \leq W \leq 1$$

Donc W prend ses valeurs dans l'ensemble [0,1]

Test de Shapiro-Wilk sous H_0

Proposition

Sous H_0, W est libre et sa loi est appellée loi de Shapiro-Wilk de paramètre n.
Lorsque n tend vers ∞ :

- *Sous H_0 W converge vers 1 en Probabilité*
- *Sous H_1, W converge vers un réel $\rho^2 < 1$ en probabilité, qui dépend de la véritable loi*

Démonstration de la liberté sous H_0

Sous H_0, pour tout i on a $Y_i = \sigma X_i + \mu$

Alors

$$W = \frac{(\sum_{i\leq n} a_i(\sigma X_{(i)} + \mu))^2}{\sum_{i\leq n}(\sigma X_i + \mu - \sigma \bar{X}_n - \mu)^2} = \frac{(\sigma \sum_{i\leq n} a_i X_{(i)} + \mu \sum_{i\leq n} a_i)^2}{\sigma^2 \sum_{i\leq n}(X_i - \bar{X}_n)^2}$$

Or

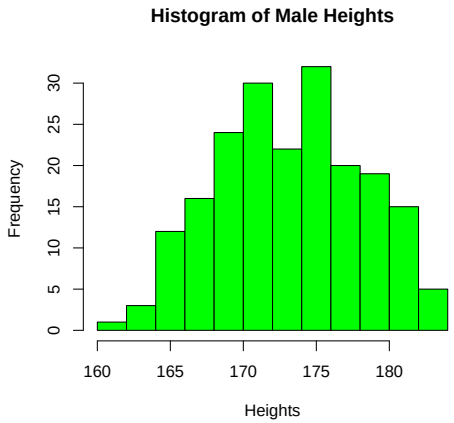
$$\sum_{i\leq n} a_i = \sum_{i\leq n} \frac{v_i}{\|v\|} = 0$$

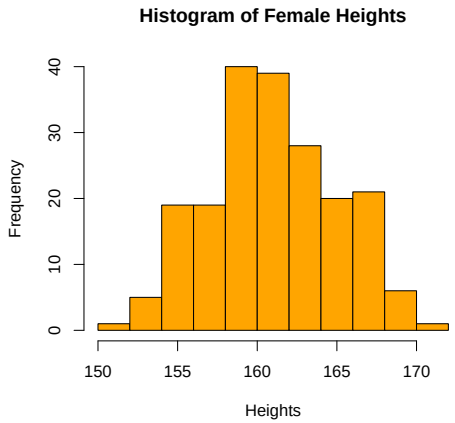
donc

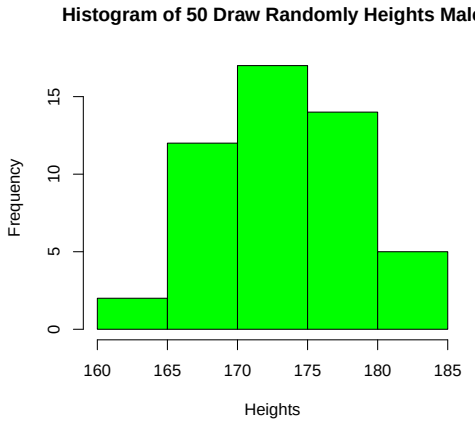
$$W = \frac{(\sum_{i\leq n} a_i X_{(i)})^2}{\sum_{i\leq n}(X_i - \bar{X}_n)^2}$$

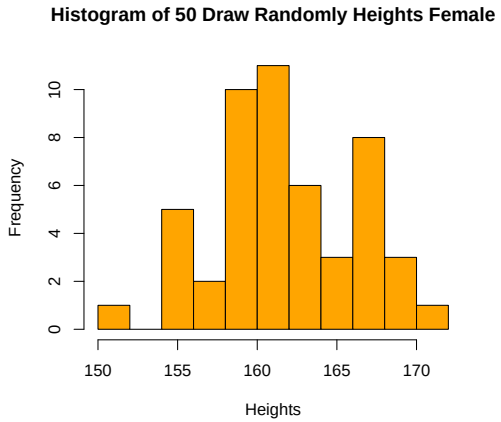
Population mondiale

On considère les quatre jeux de données suivants :









Dans R shapiro.test donne la valeur de la statistique W et la p_{value}.
On obtient

- Hommes W = 0.98778, p-value = 0.08523
- Femmes W = 0.9906, p-value = 0.221
- Hommes aléatoire W = 0.98492, p-value = 0.479
- Femmes aléatoire W = 0.9817, p-value = 0.425

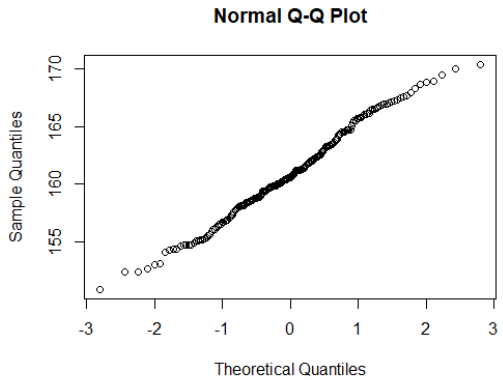


Figure 1: qqnorm sur female

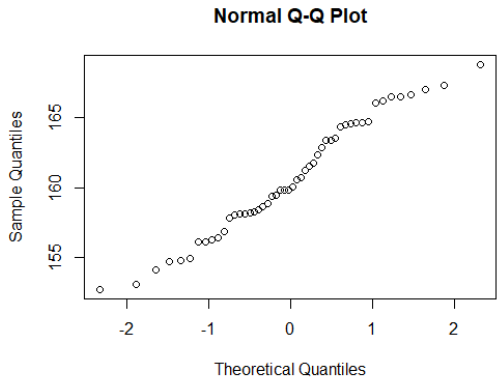


Figure 2: qqnorm sur random female

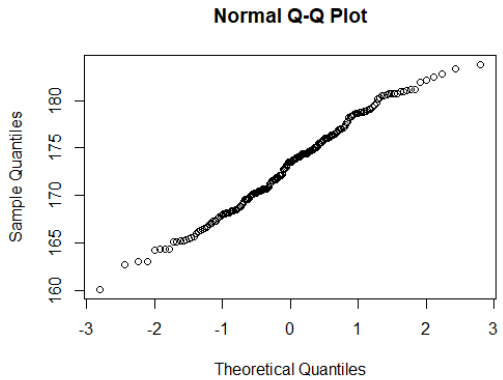


Figure 3: qqnorm sur male

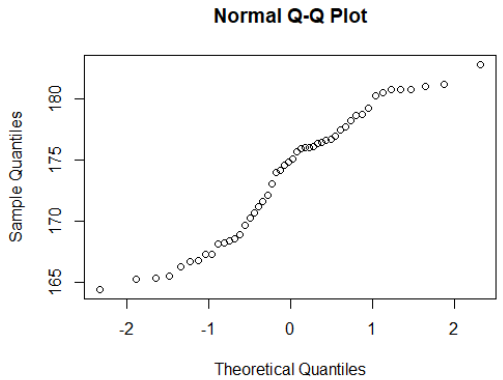


Figure 4: qqnorm sur random male

Cadre

On se place dans le cadre suivant :

- On note $Y_1, ..., Y_n$ nos observations iid
- On donne les hypothèses suivantes : $H_0 = \{P \in \mathcal{P}_0\}$ $H_1 = \{P \notin \mathcal{P}_0\}$ où $\mathcal{P}_0 = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$
- On définit $\hat{F}_n(x) = \frac{1}{n} \sum_{1 \leq n} 1_{\{Y_i \leq x\}}$
- $\xi_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_{0,\bar{x}_n,V_n}(x)|$
- $Z_n = \sup_{u \in (0,1)} |\frac{1}{n} \sum_{1 \leq n} 1_{\{U_i \leq u\}} - \psi_n(u, U_1, ..., U_n)|$
- Avec U_i va uniformes iid

Comparaison

- Première Idée : Calculer les espérances des p-value des deux tests, du vecteur aléatoire $(Y_1, .., Y_n)$, Y_i iid avec Y_1 suivant $\mathcal{N}(0,1)$. Le test ayant la plus grande valeur est meilleur.
- Deuxième Idée : Calculer la puissance des deux tests pour des lois différentes de la loi gaussienne.

1ère Idée

```
for(i in 1:100000){
  Normal = rnorm(1000, 0, 1)

  p1 = shapiro.test(Normal)$p.value
  p2 = lillie.test(Normal)$p.value
  P1 = P1 +p1
  P2 = P2 +p2
}

P1 = P1/100000
P2 = P2/100000
```

On obtient : $P1 = 0.4917$; $P2 = 0.4846$

2eme idée

La puissance de test correspond à 1 - erreur de type 2 ($\theta \in H_1$, $\mathbb{P}_\theta(X_n \notin W_n)$) . C'est à dire la probabilité sous H1 de rejeter H0.

Alternative Distribution	Sample Size (n)	Power of Test $\alpha = 0.05$ SW	LF
Gamma (4,5)	10	0.1407	0.1065
	20	0.2864	0.1771
	30	0.4442	0.2545
	50	0.6946	0.3991
	100	0.9566	0.7008
	200	0.9997	0.9518
	300	1.0000	0.9929
	400	1.0000	0.9998
	500	1.0000	1.0000
	1000	1.0000	1.0000
	2000	1.0000	1.0000
$\chi^2(4)$	10	0.2445	0.1680
	20	0.5262	0.3184
	30	0.7487	0.4650
	50	0.9484	0.6841
	100	0.9997	0.9470
	200	1.0000	0.9997
	300	1.0000	1.0000
	400	1.0000	1.0000
	500	1.0000	1.0000
	1000	1.0000	1.0000
	2000	1.0000	1.0000

Source : Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests de Adiah Mohd Razali et Bee Wah Yap.

Thank You