Part III

Les tests de normalité

7 Les différents tests

Les lois normales (ou de Gauss) sont des lois de probabilités aux propriétés remarquables. Nous aurons l'occasion, dans cette unité, de voir leur omniprésence dans les hypothèse (au sens mathématiques) des modèles statistiques.

On va donc se demander si la série est normalement distribuée. Pour valider ou infirmer ce caractère normal d'une distribution, on pourra utiliser un test de normalité.

7.1 Hypothèse nulle et puissance du test

Tous les tests de normalité classiques testent l'hypothèse nulle $\mathcal{H}_0 =$"la distribution est normale".

7.2 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et il est improbable d'obtenir de telles données en supposant qu'elles soient normalement distribues.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. Rien ne s'oppose au fait que la série soit normale (Pour autant, rien ne l'assure non plus ! D'où la nécessité de la puissance élévée pour le test).

Puissance d'un test (en bref). Si on teste $\mathcal{H}_0$ contre $\mathcal{H}_1$ unique hypothèse alternative, puissance = $P(\underbrace{\text{accepter } \mathcal{H}_1}_{\text{rejeter } \mathcal{H}_0} \mid \mathcal{H}_1 \text{ vraie})$ $= 1 - \beta$ erreur de 2ème espèce.

7.3 Les différents tests de normalité

Il existe différents tests de normalité. Certains testeront l'adéquation des fonctions de répartitions quand d'autres s'intéresseront à d'autres propriétés de la loi normale.

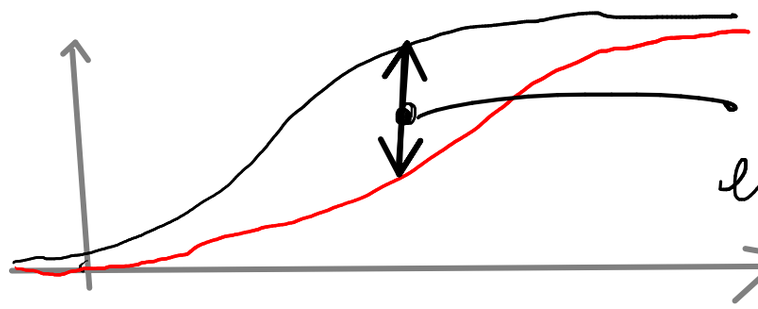
7.3.1 Le test du χ^2 d'adéquation à une loi normale

Le reproche fait à l'utilisation de ce test est la nécessité de regrouper les observations en classes. Ainsi, l'intervention de ce choix plus ou moins arbitraire lui fait perdre de la puissance.

7.3.2 Tests de Lilliefors et d'Anderson–Darling,

Ces test sont des mises en application dans le cadre normal du test de Komogorov–Smirnov qui, de manière générale s'intéresse à l'adéquation des fonctions de répartition de la loi statistique et de celle de la loi théorique d'intérêt. Ces tests ont donc pour statistique de test la distance verticale maximale entre les deux fonctions de répartition.

Le test de Lilliefors est l'application directe du test de Komogorov–Smirnov au cadre Gaussien. Ainsi, il capte peu les différences dans les queues, c'est-à-dire, l'occurrence d'événement rare.



Pb : les queues de distribution se ressemblent toujours.

On la non-normalité peut se traduire par des événements rares.

→ pénalise les écarts de fonctions de répartition dans les queues.

Le test d'Anderson–Darling tente de capter ces événements rares en pénalisant la différence qu'ils engendrent sur les fonctions de répartition. Ce test semble assez puissant, néanmoins, il reste moins populaire que le test de Shapiro–Wilk (à venir).

7.4 Tests d'Agostino et de Jarque–Bera

Ces test s'intéressent à l'aplatissement des queues et à la symmétrie des distribution, les principaux attribus des lois normales. Leur statistiques de test pénalisent donc les écarts à l'aplatissement des queues, ainsi que les écarts à la symétrie des observations. Ces tests sont relativement puissants et, ce qui est notable, le restent pour des effectifs n grands. On leur reprochera de ne pas exhiber le défaut de distribution amenant à la non normalité. Ils restent moins populaires que le test de Shapiro–Wilk.

8 Le test de Shapiro–Wilk

Le test de Shapiro–Wilk reste le plus populaire des tests de normalité, par sa puissance, mais peut-être pour le fait qu'il s'appuie sur un outil graphique de normalité qui va nous permettre de comprendre ce qui rend notre variable non normale. Il s'agit du QQ-plot, graphe quantile-quantile.

Rappel : quantile q_α : $P(X \leq q_\alpha) = \alpha$. $F_X(q_\alpha)$

$q_{0,25}$ (1er quantile) représente $P(X \leq q_{0,25}) = 0,25$

Si $F_X(q_\alpha) = \alpha$ alors $q_\alpha = F_X^{-1}(\alpha)$.

8.1 Le QQ-plot

Soit $x_1, \cdots, x_n$ une série statistique. On peut chercher à savoir si la distribution des données est gaussienne ou Poisson etc... Notons F_0 la fonction de répartition de cette loi de probabilité d'intérêt.

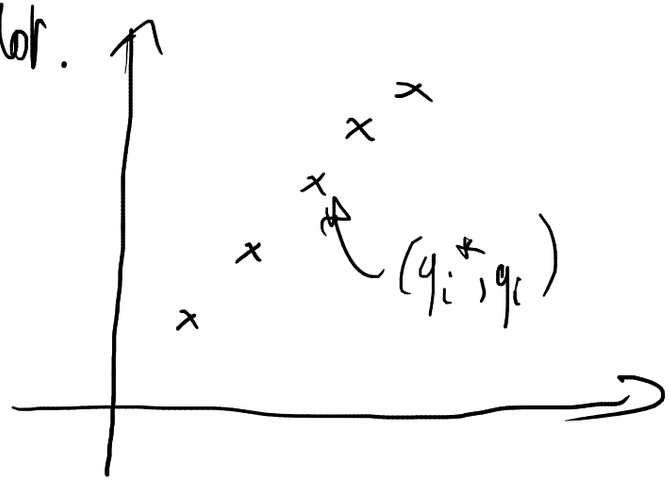
Le QQ-plot est un outil graphique permettant de visualiser rapidement l'adéquation de la distribution d'une série numérique à une distribution de référence. Dans notre contexte, on considérera une loi normale dont les paramètres seront estimés sur la série statistique observée.

Dans ce graphe, on reporte sur l'axe des ordonnées les fractiles q_i correspondant à la distribution observée et sur l'axe des abscisses ceux correspondant à la distribution théorique q_i^*. On reporte dans un graphique le nuage de points $(q_i^*; q_i)_i$.

Idée : deux distributions sont égales si tous leurs quantiles sont égaux.

Les quantiles observés q_i sont à comparer aux quantiles de la loi théorique q_i^*

QQ-plot.



On trace le nuage de points $(q_i^*; q_i)$. S'ils sont alignés sur la diagonale, il y a adéquation.

8.1.1 Données brutes

Il est primordial de classer dans l'ordre croissant les observations statistiques :

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

$P(X \leq x_{(1)}) = 1/n$ $\Rightarrow$ $q_1^* = q_{1/n} = x_{(1)}$

$P(X \leq x_{(2)}) = 2/n$ $\Rightarrow$ $q_2^* = q_{2/n} = x_{(2)}$

ordonner les

La série brute ordonnée correspond aux quantiles observés d'ordre i/n

tri obtenu sur `R` avec la fonction `sort()`. Clairement, l'observation $x_{(i)}$ est le $i$ème quantile d'ordre $1/n$, plus précisément, $q_i = x_{(i)}$ est le quantile d'ordre i/n. Il reste alors à calculer la série de quantiles théoriques

$$q_i^* = F_0^{-1}\left(\frac{i}{n}\right),$$

qu'on obtient sur `R`, dans le cas d'une loi normale, par `qnorm(i/n,mu,sigma)`. Il reste à tracer le nuage de points $(q_i^*; q_i)_i$.

Problème avec le quantile d'ordre 1 d'une loi normale : $q_1^{cP} = +\infty$

Note : Il se peut que la statistique d'ordre $x_{(i)}$ soit considérée comme le quantile d'ordre corrigé :

- $i/(n+1)$ si la population est divisée en $n+1$ tranche,
- $(2i-1)/(2n)$ si on souhaite prendre le milieu de la tranche,
- $(i-0.375)/(n+0.25)$ par certains auteurs (Saporta, 2006, p. 361).

Cela ne devrait pas changer fondamentalement les choses.

8.1.2 Données discrètes ordonnées

Dans le cas d'une variable quantitative dont les valeurs sont regroupées par modalités. Soient $m_1, \cdots, m_J$ les modalitées de la série $x_1, \cdots, x_n$ que l'on appellera quantiles observés ($q_i = m_i$). On a alors

Modalités ordonnées, quantiles observés	$q_1 = m_1$	$\cdots$	$q_J = m_J$
Fréquences cumulées	F_1	$\cdots$	F_J
Quantiles théoriques	$q_1^* = F_0^{-1}(F_1)$	$\cdots$	$q_J^* = F_0^{-1}(F_J)$

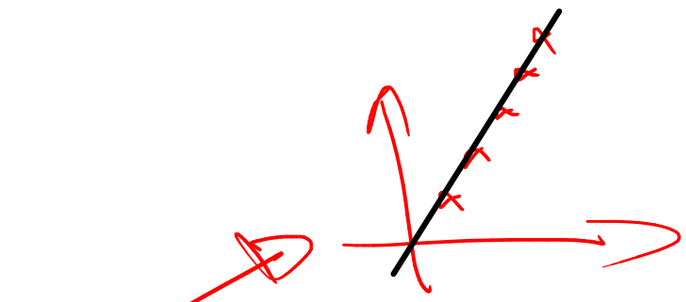
Il reste à tracer le nuage de points $(q_i^*; q_i)_i$.

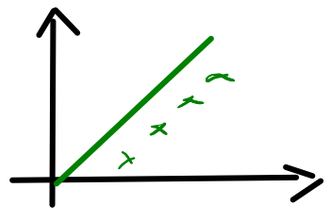
8.1.3 Données continues regroupées par classes

Dans le cas d'une variable quantitative dont les valeurs sont regroupées en classes de modalité :

Classe	$[b_0, b_1[$	$[b_1, b_2[$	$\dots$	$[b_{J-1}, b_J[$
quantiles observés(centre de la classe)	q_1	q_2	$\dots$	q_J
Fréq. cumulées	F_1	F_2	$\dots$	F_J
Quantiles théoriques	$q_1^* = F_0^{-1}(F_1)$	$q_2^* = F_0^{-1}(F_2)$	$\dots$	$q_J^* = F_0^{-1}(F_J)$

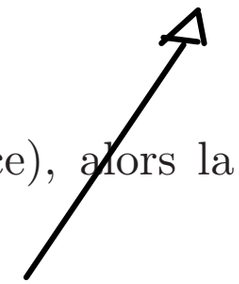
Il reste à tracer le nuage de points $(q_i^*; q_i)_i$.



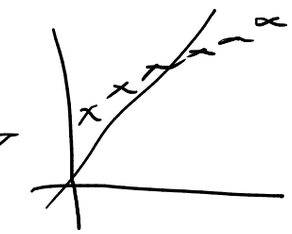


8.1.4 Interprétation

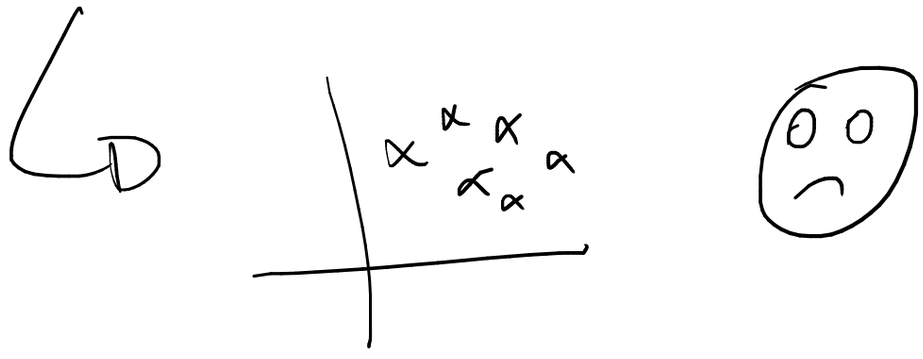
- Si les points sont alignées sur la diagonale du carré de côté 1 (première bissectrice), alors la loi théorique proposée (de fonction de répartition F_0) est adaptée aux observations.



- Si les points sont alignés sur une droite parallèle à la diagonale du carré de côté 1, on soupçonnera une erreur sur les paramètres de position de la loi théorique.



- Si les points sont alignés sur une droite passant par l'origine mais inclinée par rapport à la diagonale du carré de côté 1, on soupçonnera une erreur sur les paramètres de dispersion de la loi théorique.
- Si les points sont alignés sur une droite ne passant pas par l'origine et inclinée par rapport à la diagonale du carré de côté 1, on soupçonnera une erreur sur les paramètres de dispersion et de position de la loi théorique.
- Si les points ne sont pas alignés sur une droite, la loi théorique n'est pas adaptée aux observations.



Note : pour observer la normalité, on trace un qq-plot par rapport à une loi théorique normale. Dès l'instant où on obtient une droite, même si elle n'est pas proche de la diagonale, on a adéquation à une normale (même s'il y a problème de moyenne ou d'écart-type).

8.2 Le test de Shapiro Wilk

Le test de Shapiro–Wilk va tester l'alignement des points du qq-plot par rapport à une loi normale.

↳ "calcule un coefficient de corrélation linéaire corrigé".

Sous l'hypothèse nulle $\mathcal{H}_0$ =“la série statistique est normalement distribuée”, la statistique du test W est un coefficient de détermination corrigé[3]du qq-plot. Ainsi, $0 \leq W \leq 1$ et plus W est élevé, plus la compatibilité avec la loi normale est crédible. La région critique, rejet de la normalité, s'écrit :

$$W < W_{crit}.$$

n \ α	$0,05$	$0,01$
3	0,767	0,753
4	0,748	0,687
5	0,762	0,686
6	0,788	0,713
7	0,803	0,730
8	0,818	0,749
9	0,829	0,764
10	0,842	0,781
11	0,850	0,792
12	0,859	0,805
13	0,856	0,814
14	0,874	0,825
15	0,881	0,835
16	0,837	0,844
17	0,892	0,851
18	0,897	0,858
19	0,901	0,863
20	0,905	0,868
21	0,908	0,873
22	0,911	0,878
23	0,914	0,881
24	0,916	0,884
25	0,918	0,888
26	0,920	0,891

n \ α	$0,05$	$0,01$
27	0,923	0,894
28	0,924	0,896
29	0,926	0,898
30	0,927	0,900
31	0,929	0,902
32	0,930	0,904
33	0,931	0,906
34	0,933	0,908
35	0,934	0,910
36	0,935	0,912
37	0,936	0,914
38	0,938	0,916
39	0,939	0,917
40	0,940	0,919
41	0,941	0,920
42	0,942	0,922
43	0,943	0,923
44	0,944	0,924
45	0,945	0,926
46	0,945	0,927
47	0,946	0,928
48	0,947	0,929
49	0,947	0,929
50	0,947	0,930

8.3 Exemple

On observe la richesse des régions françaises en 2019.

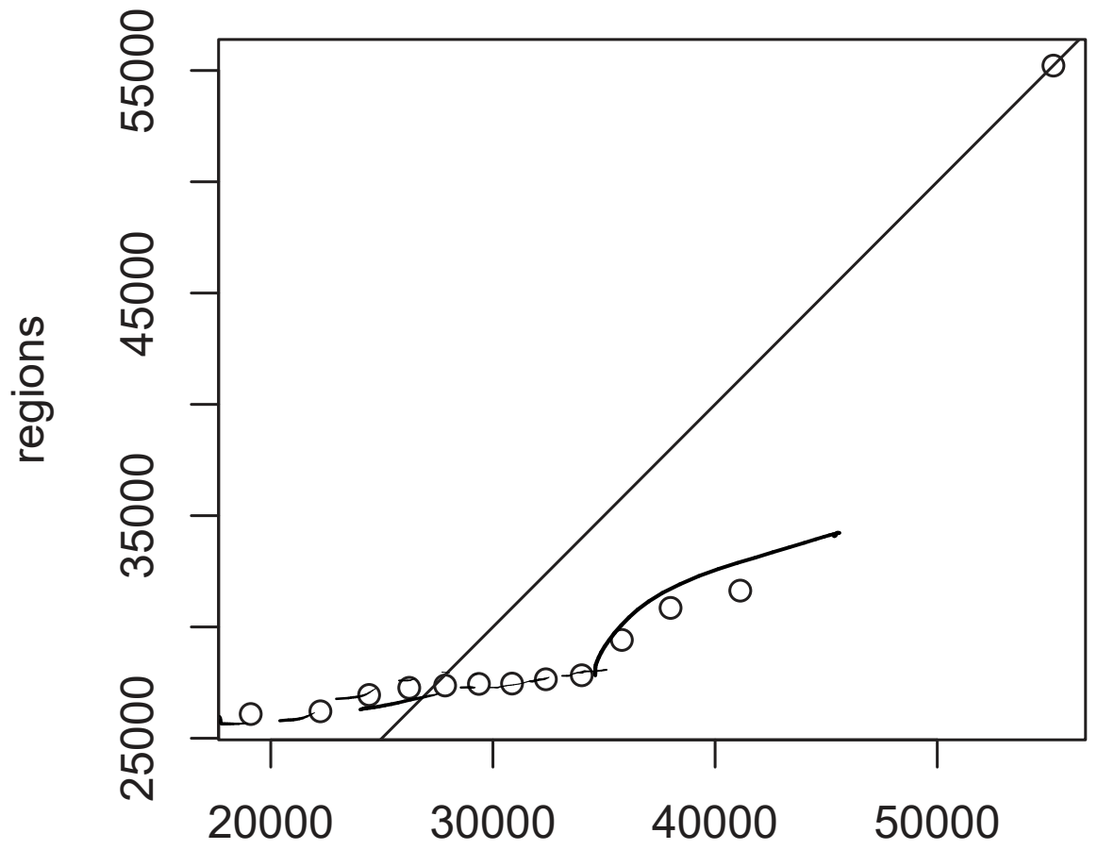
	PIB/Hab.
Auvergne Rhônes Alpes	31639
Bourgogne Franche Comté	26218
Bretagne	27838
Centre Val de Loire	27274
Corse	26954
Grand Est	27378
Hauts de France	26095
Ile de France	55227
Normandie	27465
Nouvelle Aquitaine	27657
Occitanie	27449
Pays de la Loire	29424
Provence Alpes Côte d'Azur	30864

On souhaite observer l'ajustement à une loi normale. On trie les données dans l'ordre croissant, on donne les fréquences cumulées :

x	26095	26218	26954	27274	27378	27449	27465	27657	27838	29424	30864	31639	55227
$i/13$	0.0769	0.1538	0.2308	0.3077	0.3846	0.4615	0.5385	0.6154	0.6923	0.7692	0.8462	0.9231	1

On calcule les quantiles théoriques d'après une loi normale $\mathcal{N}(30114; 7726)$ avec la fonction `qnorm(Fi, 30114,7726)`:

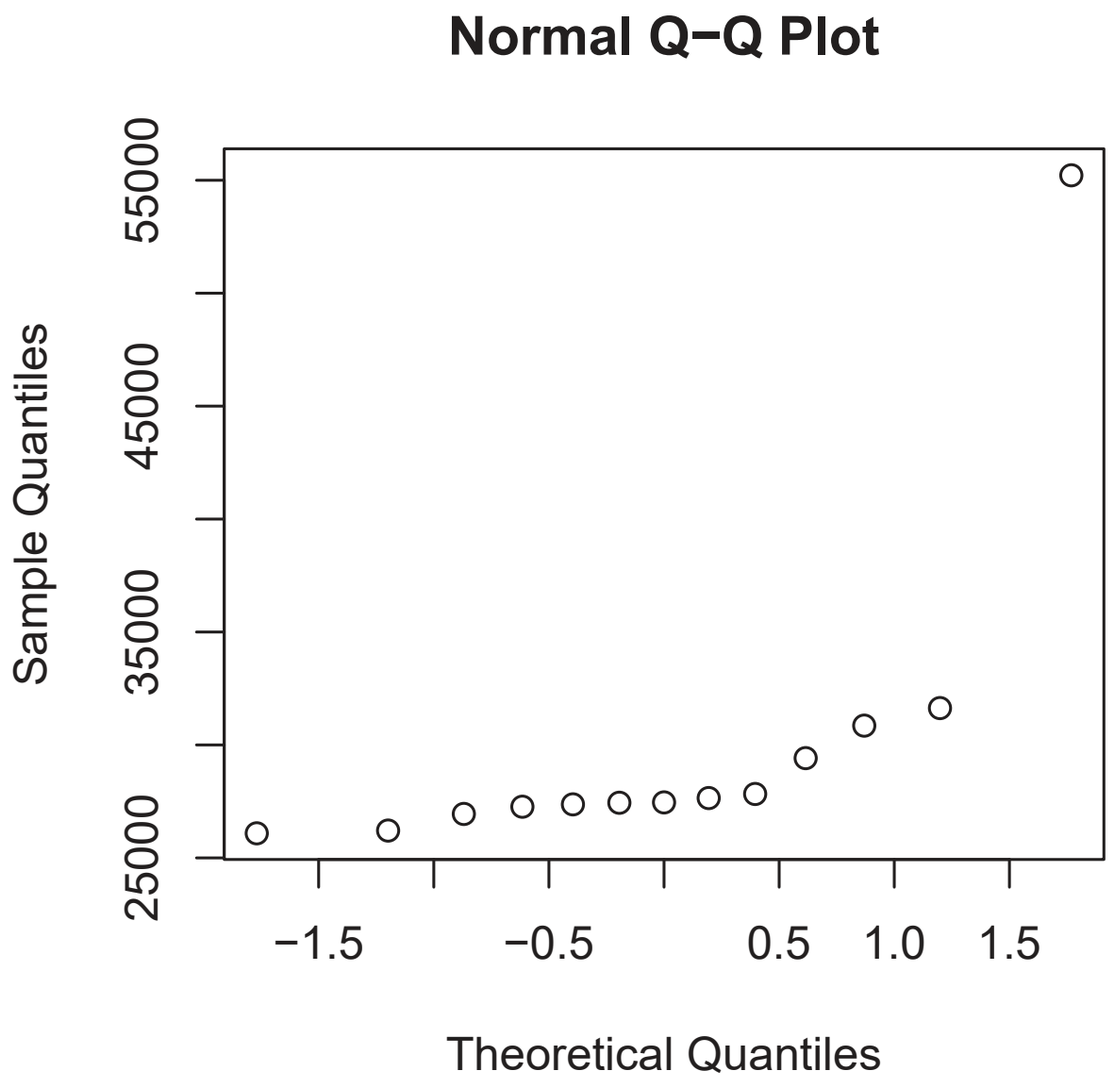
x	26095	26218	26954	27274	27378	27449	27465	27657	27838	29424	30864	31639	55227
$i/13$	0.0769	0.1538	0.2308	0.3077	0.3846	0.4615	0.5385	0.6154	0.6923	0.7692	0.8462	0.9231	1
q_i^*	19096	22232	24425	26232	27847	29367	30860	32380	33995	35802	37995	41131	Inf



Nuage non rectiligne donc

@a -plot.

On pense que ce n'est pas une loi normale.


Normal Q−Q Plot
Sample Quantiles
55000
45000
35000
25000
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

On conclut avec le test de Shapiro–Wilk.

"Normalité"

```
        Shapiro-Wilk normality test

data:  regions
W = 0.4926, p-value = 8.91e-06
```

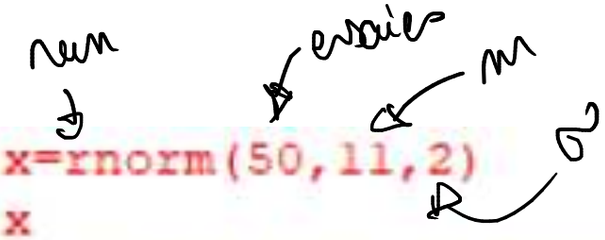
On rejette la normalité.

$p < 5\%$. On rejette H_0

La variable n'est pas normale (significativement).

8.4 Mise en œuvre du test de Shapiro–Wilk sur R

Simulons des observations normales et appliquons le test de Shapiro–Wilk.




```
> x=rnorm(50,11,2)
> x
 [1] 10.214512  9.081032  6.620966  9.362181  8.986913 11.673092 13.729038
 [8]  9.677063 11.311007 12.167166 15.143072 10.128579  9.741033  9.510930
[15] 11.741705 11.539940  9.512374 13.226112 11.564665 13.886239  7.412722
[22] 11.397585 11.934305 12.439496  7.633424 11.339367 11.585285  5.934555
[29]  9.705902 10.930072  7.706895 11.698624 13.606455 12.948415 12.573321
[36] 10.926673  7.010748  9.134820 10.628741 10.815647  8.714866  7.589055
[43] 12.306357  9.384695 10.401019 13.733044 11.899843  9.970912 14.845173
[50] 14.696913
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.98553, p-value = 0.7942
```

Non rejet de H_0 $\Rightarrow$ on peut considérer que la série est normale.

Simulons des observations uniformes et appliquons le test de Shapiro–Wilk.

```
> y=runif(50,-5,15)
> y
 [1] -4.0879249 -3.1266726 -0.7170668 -2.2280176 -1.7824772 14.2489004
 [7]  2.1059444 -3.3355710 -2.7841404  1.3467159  1.0099143  7.3365456
[13]  7.8388273 -0.3247354  1.2960432  8.6903540 10.2471285  1.0189554
[19]  3.9594622  7.2452501 12.5912116  3.7068249  1.8631227  8.8284095
[25]  0.7586813  0.9331361 12.0389597 -2.1943549  1.0193574  9.9336526
[31] -3.5596714  8.6926976  9.7736992  1.4294532 -4.8870104 11.9398889
[37] 10.3700827  3.0211665  0.5607261 -2.7961990 -2.4244256  3.3730912
[43]  0.4950279  9.9933305  3.8206514  1.2576399  7.2472824 -1.5863573
[49]  3.6153996  4.0258651
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.94087, p-value = 0.01459
```

↳ rejet de H_0 ⇒ non normalité.