Ricco Rakotomalala

Tests de normalité

Techniques empiriques et tests statistiques

Version 2.0

Université Lumière Lyon 2

Avant-propos

Ce support décrit les techniques statistiques destinées à examiner la compatibilité d'une distribution empirique avec la loi normale. On parle également de test d'adéquation à la loi normale.

Ce support se veut avant tout opérationnel. Il se concentre sur les principales formules et leur mise en oeuvre pratique avec un tableur. Autant que possible nous ferons le parallèle avec les résultats fournis par les logiciels de statistique. Le bien-fondé des tests, la pertinence des hypothèses à opposer sont peu ou prou discutées. Nous invitons le lecteur désireux d'approfondir les bases de la statistique inférentielle, en particulier la théorie des tests, à consulter les ouvrages énumérés dans la bibliographie.

Un document ne vient jamais du néant. Pour élaborer ce support, je me suis appuyé sur différentes références, des ouvrages disais-je plus tôt, mais aussi des ressources en ligne qui sont de plus en plus présents aujourd'hui dans la diffusion de la connaissance.

Les seuls bémols par rapport à ces documents en ligne sont le doute que l'on pourrait émettre sur l'exactitude des informations prodiguées, mais la plupart de leurs auteurs sont des enseignants-chercheurs qui font sérieusement leur travail ; une disponibilité plus ou moins aléatoire, au gré des migrations des serveurs et de la volonté de leurs auteurs, auquel il est très difficile de remédier ; les informations sont disparates, avec une absence d'organisation, à la différence des ouvrages qui suivent une ligne pédagogique très structurante.

Néanmoins, ces ressources en ligne renouvellent profondément le panorama des documents disponibles pour les enseignements. La gratuité n'est pas le moindre de leurs atouts.

Concernant ce document, rendons à César ce qui est à César, il a été en grande partie inspiré du manuel *Engineering Statistics Handbook* du NIST, disponible en ligne `http://www.itl.nist.gov/div898/handbook/`, notamment la section 1.3.5 *Quantitative Techniques – Distributional Measures* (`http://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm`).

Enfin, selon l'expression consacrée, ce support n'engage que son auteur. Toutes suggestions ou commentaires qui peuvent en améliorer le contenu sont le bienvenu.

Table des matières

Partie I Conformité à la loi normale

1 Techniques empiriques et méthodes graphiques 5

1.1 Histogramme de fréquence 5

1.2 Boîte à moustache 6

1.3 Coefficient d'asymétrie et d'aplatissement 7

1.4 Autres indicateurs 9

1.5 Q-Q Plot et Droite de Henry 9

2 Tests statistiques 13

2.1 Test de Shapiro-Wilk 14

2.2 Test de Lilliefors 16

2.3 Test de Anderson-Darling 18

2.4 Test de D'Agostino 21

2.5 Test de Jarque-Bera 24

2.6 Conclusion sur les tests de normalité 26

3 Tests de symétrie 29

3.1 Test de symétrie basé sur le coefficient d'asymétrie 29

3.2 Test de symétrie - Test de Wilcoxon 31

3.3 Test de symétrie - Test de Van der Waerden 35

3.4 Conclusion sur les tests de symétrie 36

4 Transformation de Box-Cox 37

4.1 Fonctions de transformation de Box-Cox 37

4.2 Approche graphique : utiliser la Droite de Henry 38

4.2.1 La droite de Henry 38

4.2.2 Exploiter la droite de Henry 39

4.2.3 Box-Cox Normality Plot 39

4.2.4 Tester la normalité 41

4.3 Approche numérique : la maximisation de la vraisemblance 42

4.3.1 Fonction de densité des variables Y et X 42

4.3.2 Expression de la fonction à optimiser / λ 43

4.3.3 Application numérique 44

A Gestion des versions 47

B Mise en oeuvre des tests de normalité dans TANAGRA 49

C Code source et packages R pour les tests de normalité 51

Littérature 53

Partie I

Conformité à la loi normale

Test d'adéquation

Un test d'adéquation permet de statuer sur la compatibilité d'une distribution observée avec une distribution théorique associée à une loi de probabilité. Il s'agit de modélisation. Nous résumons une information brute, une série d'observations, à l'aide d'une fonction analytique paramétrée. L'estimation des valeurs des paramètres est souvent un préalable au test de conformité.

Au delà de la simplification, ce test permet de valider une appréhension du processus de formation des données, il permet de savoir si notre perception du réel est compatible avec ce que nous observons. Prenons l'exemple simple du jeu de dé. A priori, nous savons que chacune des faces du dé a la même probabilité d'apparaître, tout naturellement nous pensons à une modélisation avec une loi multinomiale. Si, coincé par un margoulin dans une arrière salle d'un infâme bouge, vous constatez qu'après un bon nombre de jets, certaines valeurs ont tendance à sortir plus souvent que d'autres, il faut peut être se poser la question de la loyauté du jeu : les observations ne sont plus compatibles avec la loi théorique qui devrait générer les données.

Parmi les tests d'adéquation, la conformité à la loi normale (loi gaussienne, loi de Laplace-Gauss) revêt une importance supplémentaire. En effet, l'hypothèse de normalité des distributions sous-tend souvent de nombreux tests paramétriques (ex. comparaison de moyennes, résidus de la régression, etc.). En toute rigueur, s'assurer au préalable la compatibilité des distributions avec l'hypothèse de normalité avant de procéder au test statistique proprement dit devrait être incontournable, surtout pour les petits effectifs. Fort heureusement, ce n'est pas une contrainte forte en pratique. En effet, grâce à la notion de *robustesse*, un test peut s'appliquer même si l'on s'écarte *légèrement* des conditions d'applications initiales. Dans ce point de vue, nous pouvons dès lors nous contenter de techniques simples (ex. statistique descriptives, techniques graphiques) pour vérifier si la distribution des données est réellement inconciliable avec la distribution normale (ex. asymétrie forte, distribution avec plusieurs modes, etc.).

Dans ce support, nous présenterons dans un premier temps les techniques descriptives, notamment le très populaire graphique *Q-Q plot*. Dans un second temps, nous détaillerons plusieurs tests statistiques reconnus et implémentés dans la plupart des logiciels de statistique. Et enfin, dans un troisième temps, nous étudierons les tests de symétrie des distributions qui, à certains égards, peuvent être considérés comme des cas particuliers des tests de normalité.

Notations

Pour une population Ω donnée, nous voulons étudier la conformité de la distribution d'une v.a. continue X avec la loi normale. Nous disposons pour cela de n observations x_i.

Pour certaines techniques, nous pouvons être amenés à trier les données. Nous obtenons une série triée de manière ascendante que nous noterons $x_{(i)}$: $x_{(1)}$ correspond à la plus petite valeur observée c.-à-d. $x_{(1)} = x_{min}$, $x_{(2)}$ est la 2-ème plus petite valeur, etc.

Données

Dans ce support, nous utiliserons un fichier recensant le logarithme de l'indice de masse corporelle (IMC, Body Mass Index en anglais) de 30 personnes (Figure 0.1). Nous pouvons considérer qu'il s'agit d'un petit effectif, inapproprié pour certains tests (ex. Jarque-Bera), mais adéquat pour des visées pédagogiques : le lecteur doit pouvoir facilement reproduire les calculs [1].

LN_BODYMASS														
3.4995	3.5381	3.1398	3.8979	3.4935	3.4812	3.5723	3.5056	3.5582	3.6055	3.2027	3.6055	3.3776	3.2884	3.1091
3.1135	3.3911	3.5056	3.1311	3.3945	3.4404	3.4144	4.0843	3.1864	3.1781	3.4935	3.2229	3.7705	3.4177	3.4657

Fig. 0.1. Données initiales

Nous utiliserons principalement le tableur EXCEL dans ce support, mais à plusieurs reprises nous ferons appel à des logiciels gratuits tels que TANAGRA et R, et des logiciels commerciaux tels que SPSS et STATISTICA.

1. Le fichier de données est accessible sur le Web, `http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`

1

Techniques empiriques et méthodes graphiques

L'appréhension d'un jeu de données passe systématiquement par les statistiques descriptives. Elles donnent une image globale. Bien souvent, elles permettent de se faire une idée sur les techniques que l'on pourrait utiliser et les dangers ou artefacts dont il faudra se méfier.

Bien avant les techniques complexes et les ratios savants, quelques indicateurs usuels et des graphiques judicieusement choisis sont le bienvenu. Ces outils sont disponibles dans tous les outils de traitement exploratoire des données.

1.1 Histogramme de fréquence

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

Dans la plupart des logiciels, le nombre k d'intervalles est défini de manière arbitraire, dans d'autres il est paramétrable. Une règle simple pour définir le bon nombre d'intervalles est d'utiliser la règle $k = \log_2(n)$.

Results

Attribute	Stats		Histogram			
	Statistics		**Values**	**Count**	**Percent**	**Histogram**
LN_BODYMASS	Average	3.4362	x_<_3.2066	7	23.33%	
	Median	3.4531	3.2066_=<_x_<_3.3041	2	6.67%	
	Std dev. [Coef of variation]	0.2290 [0.0666]	3.3041_=<_x_<_3.4016	3	10.00%	
	MAD [MAD/STDDEV]	0.1693 [0.7392]	3.4016_=<_x_<_3.4992	7	23.33%	
	Min * Max [Full range]	3.11 * 4.08 [0.98]	3.4992_=<_x_<_3.5967	6	20.00%	
	1st * 3rd quartile [Range]	3.22 * 3.54 [0.32]	3.5967_=<_x_<_3.6942	2	6.67%	
	Skewness (std-dev)	0.7476 (0.4269)	3.6942_=<_x_<_3.7917	1	3.33%	
	Kurtosis (std-dev)	1.1296 (0.8327)	3.7917_=<_x_<_3.8892	0	0.00%	
			3.8892_=<_x_<_3.9868	1	3.33%	
			x>=_3.9868	1	3.33%	

Fig. 1.1. Statistiques descriptives

Dans le résultat que nous reproduisons (Figure 1.1), la valeur $k = 10$ est manifestement trop élevée, il y a trop peu d'observations dans chaque intervalle. On peut essayer de descendre à $k = \log_2(30) \approx 5$ (Figure 1.2), mais définitivement il y a trop peu d'observations pour se donner une idée précise dans cet exemple.

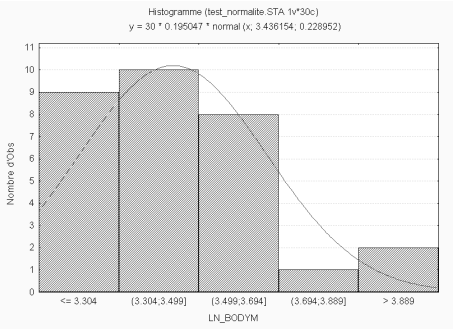


Fig. 1.2. Histogramme de fréquences

Certains logiciels procèdent automatiquement à l'estimation des deux principaux paramètres de la loi normale (μ la moyenne, σ l'écart-type) et tracent la fonction de densité correspondante pour apprécier le rapprochement entre la distribution empirique (histogramme) et la distribution théorique (Figure 1.2).

La moyenne est estimée à l'aide de la moyenne empirique :

$$\bar{x} = \frac{1}{n}\sum_i x_i = 3.4362 \tag{1.1}$$

On utilise l'estimateur non biaisé de l'écart-type :

$$s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2} = 0.2290 \tag{1.2}$$

1.2 Boîte à moustache

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

Dans notre fichier (Figure 1.3), $M = 3.4531$, $Q_1 = 3.2229$ et $Q_3 = 3.5381$. On constate un certain étalement de la distribution vers les grandes valeurs, chose que l'on pouvait déjà percevoir dans l'histogramme de fréquences (Figure 1.2).

1. `http://fr.wikipedia.org/wiki/Boîte_à_moustaches` et `http://www.sfds.asso.fr/groupes/statvotre/Boite-a-moustaches.pdf`

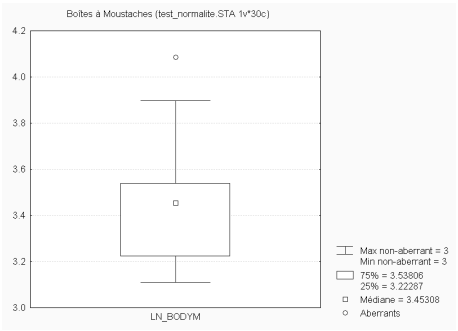


Fig. 1.3. Boîte à moustaches

Remarque 1 (Détection et influence des points atypiques). Les extrémités des moustaches sont délimités par 1.5 fois l'intervalle inter-quartile $(Q_3 - Q_1)$. Cela permet de déceler l'existence d'un point extrême[2]. Il s'agit de l'observation correspondant à $x_{max} = 4.084$, elle est largement plus élevée que les autres valeurs. Ce point est mis en évidence dans la boîte à moustaches (Figure 1.3).

Cette règle de détection est plus fiable que la fameuse règle des 3-sigma qui consiste à isoler les points en-deçà ou au-delà de 3-fois l'écart-type autour de la moyenne. En effet, elle ne repose pas sur une hypothétique symétrie de la distribution, elle utilise également des paramètres de localisation (les quartiles) qui, à la différence de la moyenne empirique, sont peu influencés par les points extrêmes.

LN_BODYMASS														
3.4995	3.5381	3.1398	3.8979	3.4935	3.4812	3.5723	3.5056	3.5582	3.6055	3.2027	3.6055	3.3776	3.2884	3.1091
3.1135	3.3911	3.5056	3.1311	3.3945	3.4404	3.4144	3.1864	3.1781	3.4935	3.2229	3.7705	3.4177	3.4657	

Fig. 1.4. Données sans le point extrême

Dans notre fichier, il est patent que la valeur 4.084 est largement plus élevée que les autres. Or tous les indicateurs et tests que nous mettrons en oeuvre reposent, au moins en partie, sur la moyenne empirique $(\bar{x})$. Il paraît plus judicieux de **supprimer cette observation**.

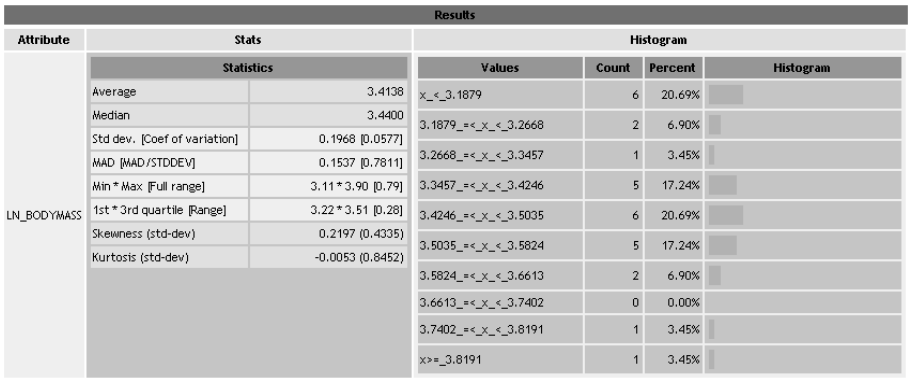
Désormais, le fichier utilisé dans les traitements comptera $n = 29$ observations (Figure 1.4), nous recalculons dès lors les statistiques descriptives (Figure 1.5).

1.3 Coefficient d'asymétrie et d'aplatissement

La loi normale est caractérisée par un coefficient d'asymétrie et un coefficient d'aplatissement nuls. Il paraît naturel de calculer ces indicateurs pour se donner une idée, ne serait-ce que très approximative, du rapprochement possible de la distribution empirique avec une gaussienne.

Plutôt que les indicateurs triviaux dérivés de la définition théorique des coefficients, les logiciels calculent les estimateurs non-biaisés.

2. `http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm`



Results

Attribute	Stats		Histogram			
	Statistics		Values	Count	Percent	Histogram
LN_BODYMASS	Average	3.4138	x_<_3.1879	6	20.69%	
	Median	3.4400	3.1879_=<_x_<_3.2668	2	6.90%	
	Std dev. [Coef of variation]	0.1968 [0.0577]	3.2668_=<_x_<_3.3457	1	3.45%	
	MAD [MAD/STDDEV]	0.1537 [0.7811]	3.3457_=<_x_<_3.4246	5	17.24%	
	Min * Max [Full range]	3.11 * 3.90 [0.79]	3.4246_=<_x_<_3.5035	6	20.69%	
	1st * 3rd quartile [Range]	3.22 * 3.51 [0.28]	3.5035_=<_x_<_3.5824	5	17.24%	
	Skewness (std-dev)	0.2197 (0.4335)	3.5824_=<_x_<_3.6613	2	6.90%	
	Kurtosis (std-dev)	-0.0053 (0.8452)	3.6613_=<_x_<_3.7402	0	0.00%	
			3.7402_=<_x_<_3.8191	1	3.45%	
			x>=_3.8191	1	3.45%	

Fig. 1.5. Statistiques descriptives sans le point extrême

Pour le coefficient d'asymétrie γ_1, appelé *skewness* en anglais, nous utilisons [3] :

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^3 = 0.2197 \tag{1.3}$$

Pour le coefficient d'aplatissement γ_2, appelé *kurtosis* en anglais, nous utilisons [4] :

$$G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} = -0.0053 \tag{1.4}$$

Si ces indicateurs sont *suffisamment proches* de la valeur 0, l'hypothèse de compatibilité avec la loi normale ne peut être rejetée. Tout le problème est de quantifier ce degré de proximité. Il faudrait connaître la loi de probabilité de ces indicateurs pour mettre en place un test statistique permettant de déterminer si l'écart est significatif ou non ; ou tout du moins, calculer les écart-type (cf. les valeurs entre parenthèses fournies par le logiciel TANAGRA, figure 1.5) et utiliser les distributions asymptotiques pour réaliser le test.

Nous détaillerons ces procédures plus loin. A ce stade, les coefficients d'asymétrie et d'aplatissement sont uniquement calculés à titre indicatif. Nous constatons néanmoins, sans trop s'avancer quant aux résultats des tests, qu'elles s'éloignent peu des valeurs de référence. L'adéquation à la loi normale paraît plausible.

Remarque 2 (Calculs avec le point atypique). Par curiosité, nous reprenons ces mêmes indicateurs en incluant le point extrême (Figure 1.1), nous constatons qu'elles prennent des valeurs sensiblement différentes, $G_1 = 0.7476$ et $G_2 = 1.1296$, confirmant, si besoin est, qu'un individu s'écartant significativement de la population peut fausser les résultats.

3. http://en.wikipedia.org/wiki/Skewness
4. http://en.wikipedia.org/wiki/Kurtosis

1.4 Autres indicateurs

D'autres indicateurs peuvent être mis à profit pour apprécier rapidement l'écart à la loi normale. Par exemple, la distribution étant symétrique, l'écart entre la médiane (M) et la moyenne empirique ($\bar{x}$) ne devrait pas être très élevé. Dans notre jeu de données, la médiane est égale à 3.4400 et la moyenne 3.4138 (Figure 1.5). Ce dispositif est toutefois très grossier : l'importance de l'écart dépend de la dispersion des données, il permet uniquement d'apprécier la symétrie de la distribution.

$$D = \frac{1}{n}\sum_i |x_i - \bar{x}| \tag{1.5}$$

Autre caractéristique d'une gaussienne, le rapport entre l'écart absolu moyen (*MAD - mean absolute deviation* en anglais, équation 1.5) et l'écart-type est asymptotiquement égal à $\sqrt{\frac{2}{\pi}} \approx 0.7979$. Dans notre fichier de données, il est de 0.7811 (Figure 1.5). Ici également, il paraît difficile de rejeter d'emblée l'adéquation à la loi normale.

Ce dispositif peut être étoffé et aboutir à un test statistique fondé sur le ratio écart absolu moyen - écart-type (Aïvazian et al., page 301). Des tables sont disponibles pour définir les régions critiques associés aux différents niveaux de risque. Mais il faut reconnaître que cette procédure est très rarement référencée. Pour ma part, je ne l'ai jamais vue implémentée dans un logiciel.

1.5 Q-Q Plot et Droite de Henry

Le Q-Q plot, *quantile-quantile plot*, est une technique graphique qui permet de comparer les distributions de deux ensembles de données [5].

Les échantillons ne sont pas forcément de même taille. Il se peut également, et c'est ce qui nous intéresse dans le cas présent, qu'un des ensembles de données soient générées à partir d'une loi de probabilité qui sert de référentiel.

Concrètement, il s'agit

1. de trier les données de manière croissante pour former la série $x_{(i)}$;
2. à chaque valeur $x_{(i)}$, nous associons la fonction de répartition empirique $F_i = \frac{i-0.375}{n+0.25}$ (Saporta, page 361) ;
3. nous calculons les quantiles successifs $z*_{(i)}$ d'ordre F_i en utilisant l'inverse de la loi normale centrée et réduite [6] ;
4. enfin, les données initiales n'étant pas centrées et réduites, nous dé-normalisons les données en appliquant la transformation $x*_{(i)} = z*_{(i)} \times s + \bar{x}$.

x(i)	F(i)	z(i)	x*(i)
3.109	0.02136752	-2.026	3.015
3.114	0.05555556	-1.593	3.100
3.131	0.08974359	-1.342	3.150
3.140	0.12393162	-1.156	3.186
3.178	0.15811966	-1.002	3.217
3.186	0.19230769	-0.869	3.243
3.203	0.22649573	-0.750	3.266
3.223	0.26068376	-0.641	3.288
3.288	0.29487179	-0.539	3.308
3.378	0.32905983	-0.443	3.327
3.391	0.36324786	-0.350	3.345
3.395	0.39743590	-0.260	3.363
3.414	0.43162393	-0.172	3.380
3.418	0.46581197	-0.086	3.397
3.440	0.50000000	0.000	3.414
3.466	0.53418803	0.086	3.431
3.481	0.56837607	0.172	3.448
3.493	0.60256410	0.260	3.465
3.493	0.63675214	0.350	3.483
3.500	0.67094017	0.443	3.501
3.506	0.70512821	0.539	3.520
3.506	0.73931624	0.641	3.540
3.538	0.77350427	0.750	3.561
3.558	0.80769231	0.869	3.585
3.572	0.84188034	1.002	3.611
3.605	0.87606838	1.156	3.641
3.605	0.91025641	1.342	3.678
3.770	0.94444444	1.593	3.727
3.898	0.97863248	2.026	3.813

Fig. 1.6. Tableau de calcul du q-q plot

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

Les calculs sont résumés dans un tableau de calcul que l'on peut construire facilement dans un tableur (Figure 1.6). Nous obtenons un graphique nuage de points, la droite de référence est matérialisée par la diagonale principale (Figure 1.7). Nous constatons que les points sont relativement alignés. Nous n'observons pas un écartement significatif, aucun point ne semble non plus se démarquer des autres.

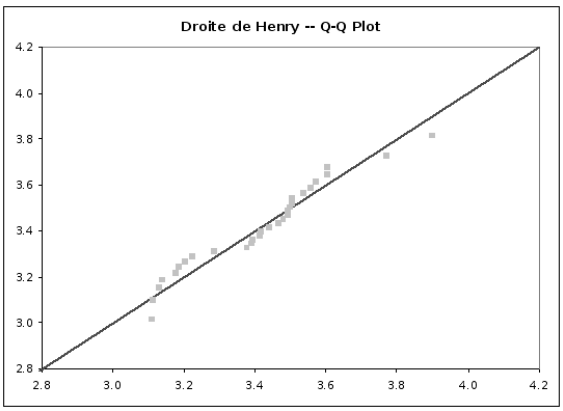


Fig. 1.7. Q-Q plot pour notre jeu de données de 29 observations

5. `http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm`
6. =LOI.NORMALE.STANDARD.INVERSE(...) dans le tableur EXCEL

Remarque 3 (Estimation de la fonction de répartition F_i). Nous n'utilisons pas l'estimation triviale $F_i = \frac{i}{n}$ dans les calculs. Il s'agit en effet de "lisser" la fonction de répartition en prenant, non pas la valeur brute, mais la valeur espérée en référence à la loi de répartition, la loi normale dans notre cas. Voir `http://www.uic.edu/classes/idsc/ids577/nscores.htm`; Blom's Normal Score - `http://www.vni.com/products/imsl/jmsl/v30/api/com/imsl/stat/Ranks.html`; quelques scores usuels pour les tests basés sur les rangs (Wilcoxon, Van der Waerden, Savage, Siegel-Tukey, Klotz, etc.) - `http://v8doc.sas.com/sashtml/stat/chap47/sect17.htm`. Plus généralement, la formule idoine est $F_i = \frac{i+a}{n+1+2a}$, $a = -\frac{3}{8} = -0.375$ est une possibilité pour la loi normale, mais d'autres variantes existent `http://en.wikipedia.org/wiki/Qq_plot`.

Remarque 4 (Variantes de la droite de Henry). D'autres modes de représentation de la droite de Henry sont couramment utilisés dans la littérature. Nous pouvons laisser les points $z*_{(i)}$ en ordonnée du graphique. L'intérêt est qu'il est possible de déterminer graphiquement les paramètres de localisation et d'échelle de la distribution empirique (par exemple, la droite coupe l'axe des abcisses à une coordonnée qui permet d'estimer μ)(Figure 1.9).

Autre représentation très populaire, nous utilisons directement en ordonnée les valeurs de F_i en utilisant un repère spécifique dit *repère gausso-arithmétique*. L'astuce est de disposer, non pas régulièrement les valeurs de la fréquence cumulée en ordonnée, mais selon une échelle qui permet d'obtenir une droite si la distribution était gaussienne[7] (Figure 1.8).

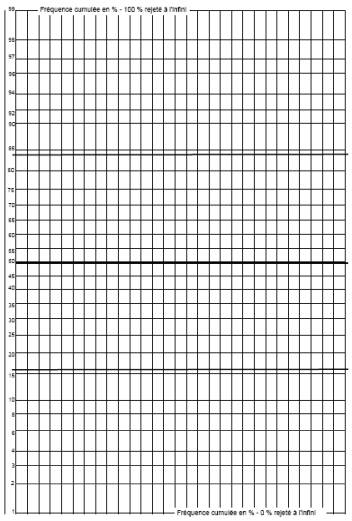


Fig. 1.8. Exemple de papier gausso-arithmétique

Remarque 5 (De l'utilisation du papier gausso-arithmétique). Ce type de papier très spécifique, vendu naguère dans les librairies, était pratique car il évitait au statisticien d'avoir à calculer à partir des tables

7. Des exemples de papier gausso-arithmétique : `http://nte-serveur.univ-lyon1.fr/nte/immediato/math2002/Tables/papier_gausso.htm`; `http://www.iut.u-bordeaux4.fr/gea/pagesweb/henry.pdf`

statistiques les valeurs successives de $z*_{(i)}$ à partir des fréquences F_i. Il n'a plus vraiment d'utilité de nos jours, un tableur fournit très facilement ces valeurs.

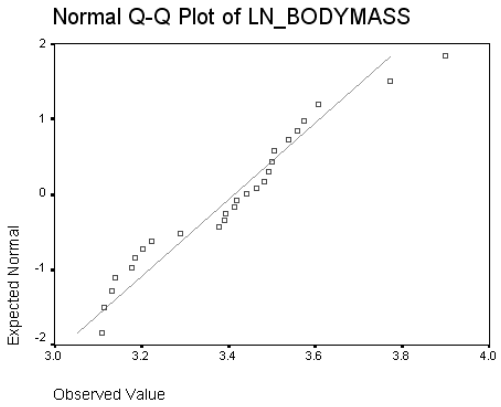


Fig. 1.9. Q-Q plot, valeurs standardisées en ordonnées

2

Tests statistiques

Très commodes, les approches empiriques n'ont pas la rigueur des techniques statistiques. Dans ce chapitre, nous présentons les tests de compatibilité à la loi normale. Encore une fois, il s'agit bien de vérifier l'adéquation (la compatibilité) à la loi normale et non pas déterminer la loi de distribution.

Mis à part le test de Shapiro-Wilk, tous les tests présentés dans ce chapitre sont, soit des variantes plus puissantes du test de Kolmogorov-Smirnov, soit basés sur les coefficients d'asymétrie et d'aplatissement.

La majorité de ces techniques sont présents dans les logiciels. Nous reprenons notamment les résultats de TANAGRA (Figure 2.1). Pour des raisons pédagogiques, nous reproduisons tous les calculs dans un tableur afin que le lecteur puisse accéder aux détail des méthodes.

A tout test est associé un risque α dit de première espèce, il s'agit de la probabilité de rejeter l'hypothèse de normalité alors qu'elle est vraie. Plus nous diminuons sa valeur, plus notre propension à accepter l'adéquation à une gaussienne est élevée. Dans tous nos exemples, nous adopterons le risque $\alpha = 5\%$.

Normality Test 1

Parameters

Attributes : 1
Examples : 29

Results

Attribute	Mu ; Sigma	Shapiro-Wilk (p-value)	Lilliefors D = max[D-,D+] (p-value)	Anderson-Darling (p-value)	d'Agostino (p-value)
X	3.4138 ; 0.1968	0.945634 (0.1408)	0.1176 = max[0.1176,0.1096] (p >= 0.20)	0.588535 (p >= 0.10)	0.5349 ^2 + 0.2259 ^2 = 0.3372 (0.8449)

Fig. 2.1. Tests de normalité avec le logiciel TANAGRA

Toutes les techniques que nous présentons dans ce chapitre sont, et ne sont que, des techniques numériques. Les résultats, rejet ou acceptation de la normalité, peuvent masquer des situations très disparates. De plus, ces tests sont très influencés par la taille de l'échantillon. La compatibilité avec la loi normale est bien (trop) souvent la règle sur des petits effectifs ; en revanche, l'incompatibilité avec la loi normale est quasi-systématiquement décidée sur de gros effectifs, même si les écarts de distributions sont faibles. De fait, les approches empiriques, notamment graphiques, gardent toute leur importance.

2.1 Test de Shapiro-Wilk

Description

Très populaire, le test de Shapiro-Wilk [1] est basé sur la statistique W. En comparaison des autres tests, il est particulièrement puissant pour les petits effectifs ($n \leq 50$). La statistique du test s'écrit :

$$W = \frac{\left[\sum_{i=1}^{[\frac{n}{2}]} a_i \left(x_{(n-i+1)} - x_{(i)}\right)\right]^2}{\sum_i (x_i - \bar{x})^2} \tag{2.1}$$

où

- $x_{(i)}$ correspond à la série des données triées ;
- $[\frac{n}{2}]$ est la partie entière du rapport $\frac{n}{2}$;
- a_i sont des constantes générées à partir de la moyenne et de la matrice de variance co-variance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont fournies dans des tables spécifiques [2]

La statistique W peut donc être interprétée comme le coefficient de détermination (le carré du coefficient de corrélation) entre la série des quantiles générées à partir de la loi normale et les quantiles empiriques obtenues à partir des données. Plus W est élevé, plus la compatibilité avec la loi normale est crédible. La région critique, rejet de la normalité, s'écrit :

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

Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk [3].

Calculs

Les calculs s'agencent de la manière suivante (Figure 2.2) :

1. trier les données x_i, nous obtenons la série $x_{(i)}$;
2. calculer les écarts $\left(x_{(n-i+1)} - x_{(i)}\right)$;
3. lire dans la table pour $n = 29$, les valeurs des coefficients a_i ;
4. former le numérateur de W, $nW = 1.0231$;
5. former le dénominateur de W, $dW = 1.0847$;
6. en déduire $W = \frac{1.0240}{1.0856} = 0.9432$;
7. pour une risque $\alpha = 0.05$, le seuil critique lue dans la table pour $n = 29$ est $W_{crit} = 0.926$.

Dans notre exemple, $W > W_{crit}$, au risque de 5%, l'hypothèse de normalité est compatible avec nos données.

1. http://www.educnet.education.fr/rnchimie/math/benichou/tests/normalite/normalite.htm
2. http://www.educnet.education.fr/rnchimie/math/benichou/tables/tshapiro/coef.htm ou http://www.santetropicale.com/SANTEMAG/algerie/stat/stat_10.htm#28
3. http://www.educnet.education.fr/rnchimie/math/benichou/tables/tshapiro/tshapiro.htm

Moyenne
3.414

	x(i)	écart	a_i	nW	dW
1	3.109	0.789	0.4291	0.3386	0.0929
2	3.114	0.656	0.2968	0.1947	0.0899
3	3.131	0.474	0.2499	0.1185	0.0800
4	3.140	0.465	0.2150	0.1000	0.0749
5	3.178	0.394	0.1854	0.0730	0.0556
6	3.186	0.372	0.1616	0.0601	0.0519
7	3.203	0.335	0.1395	0.0467	0.0444
8	3.223	0.283	0.1192	0.0337	0.0364
9	3.288	0.218	0.1002	0.0218	0.0158
10	3.378	0.122	0.0822	0.0100	0.0013
11	3.391	0.102	0.0650	0.0066	0.0005
12	3.395	0.098	0.0483	0.0047	0.0004
13	3.414	0.067	0.0320	0.0021	0.0000
14	3.418	0.048	0.0159	0.0008	0.0000
15	3.440	-	0.0000	0.0000	0.0007
16	3.466	-	-	-	0.0027
17	3.481	-	-	-	0.0045
18	3.493	-	-	-	0.0063
19	3.493	-	-	-	0.0063
20	3.500	-	-	-	0.0074
21	3.506	-	-	-	0.0085
22	3.506	-	-	-	0.0085
23	3.538	-	-	-	0.0154
24	3.558	-	-	-	0.0208
25	3.572	-	-	-	0.0250
26	3.605	-	-	-	0.0366
27	3.605	-	-	-	0.0366
28	3.770	-	-	-	0.1269
29	3.898	-	-	-	0.2345
			Somme	**1.0115**	**1.0847**

nW	1.0231
dW	1.0847

W	0.9432

Fig. 2.2. Test de Shapiro-Wilk avec un tableur

Implémentations et logiciels

Comme nous pouvons le constater, les calculs sont assez complexes et reposent sur des valeurs tabulées avec une certaine précision. Il importe de vérifier les différentes variantes implémentées dans les logiciels.

Petits effectifs

Pour les petits effectifs ($n \leq 50$), SPSS procède au calcul exact et propose la valeur $W = 0.9438$. Il n'est pas opérant en revanche dès que $n > 50$. Ce résultat est très proche de ce que nous obtenons avec le tableur. A la différence que les coefficients a_i doivent être vraisemblablement plus précis dans SPSS.

Effectifs intermédiaires

Pour les effectifs de taille modérée, un autre algorithme prend le relais. Le programme de référence a été publiée dans la revue *Applied Statistics Journal*[4], le code source FORTRAN est accessible en ligne[5]. Il donne des résultats précis jusqu'à $n \leq 5000$. Il produit aussi la probabilité critique (*p-value*) du test.

Il est implémenté dans le logiciel DATAPLOT du NIST[6]. Nous ne l'avons pas testé. En revanche, l'implémentation dans le logiciel R a été évaluée (fonction *shapiro.test(...)*[7]). Nous obtenons la valeur $W = 0.9456$, avec une p-value $= 0.1408$. L'hypothèse de normalité ne peut être rejetée.

4. Algorithm AS R94 (SWILK sub routine) from the Applied Statistics Journal, 1995, Vol. 44, No. 4.
5. http ://lib.stat.cmu.edu/apstat/R94
6. `http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/wilkshap.htm`
7. Voir le prototype de la fonction `http://sekhon.berkeley.edu/stats/html/shapiro.test.html`

Le code source en FORTRAN a été porté en DELPHI dans le logiciel TANAGRA, nous obtenons exactement les mêmes résultats (Figure 2.1).

STATISTICA, *dixit le fichier d'aide*, s'appuie sur une extension de l'algorithme de Royston (1982) [8]. Il y a de fortes chances qu'il s'agit d'une version fort similaire à celle du même auteur en 1995 (DATAPLOT). Nous obtenons également des valeurs identiques.

2.2 Test de Lilliefors

Description

Le test de Lilliefors [9] est une variante du test de Kolmogorov-Smirnov où les paramètres de la loi (μ et σ) sont estimées à partir des données. La statistique du test est calculée de la même manière. Mais sa loi est tabulée différemment, les valeurs critiques sont modifiées pour un même risque α. Elles ont été obtenues par simulation.

Les avis sont partagés quant à la puissance de ce test. Il semble qu'il soit sensible au désaccord de la distribution empirique avec la loi théorique aux alentours de la partie centrale de la distribution, là où justement les écarts ont peu d'effets sur les tests paramétriques. Il est moins performant en revanche lorsque le désaccord porte sur les queues de distribution, pourtant préjudiciables. Certains le déconseillent et préfèrent le test de Shapiro-Wilk ou les tests basés sur les coefficients d'asymétrie et d'aplatissement [10].

La statistique du test d'écrit :

$$D = \max_{i=1,\ldots,n} \left(F_i - \frac{i-1}{n}, \frac{i}{n} - F_i \right) \tag{2.2}$$

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

La table des valeurs critiques D_{crit} pour les petites valeurs de n et différentes valeurs de α doivent être utilisées [11]. Lorsque les effectifs sont élevés, typiquement $n \geq 30$, il est possible d'approcher la valeur critique à l'aide de formules simples :

α	Valeur critique D_{crit}
0.10	$\frac{0.805}{\sqrt{n}}$
0.05	$\frac{0.886}{\sqrt{n}}$
0.01	$\frac{1.031}{\sqrt{n}}$

8. Patrick Royston (1982) Algorithm AS 181 : The W Test for Normality. Applied Statistics, 31, 176Ũ180.
9. Lilliefors, H. (June 1967), "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", Journal of the American Statistical Association, Vol. 62. pp. 399-402.
10. `http://en.wikipedia.org/wiki/Lilliefors_test`
11. `http://courses.wcupa.edu/rbove/eco252/252KStest.doc`

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

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

Remarque 6 (Calcul de la p-value). Abdi et Molin (2007) fournissent des approximations plus précises [12]. Surtout, ils proposent une formule, assez complexe il faut le reconnaître, pour obtenir la probabilité critique (*p-value*) du test. Cela simplifie beaucoup la procédure, il suffit de comparer cette *p-value* avec le risque α que l'on s'est choisi. Néanmoins, je ne connais pas à l'heure actuelle de logiciel qui ait intégré cette formule.

Calculs

Le test se construit comme le test de Kolmogorov-Smirnov, à la différence que les paramètres de la loi sont estimés et que les valeurs critiques modifiés (Figure 2.3) :

1. les données sont triées pour former la série $x_{(i)}$;
2. nous estimons les paramètres, $\bar{x} = 3.4138$ et $s = 0.1968$;
3. nous calculons alors les données centrées et réduites $z_{(i)} = \frac{x_{(i)} - \bar{x}}{s}$;
4. nous utilisons la fonction de répartition de la normale centrée et réduite [13] pour obtenir les fréquences théoriques F_i ;
5. que nous opposons aux fréquences empiriques pour obtenir la statistique D du test [14], en calculant tour à tour $D^- = \max_{i=1,\ldots,n}\left(F_i - \frac{i-1}{n}\right) = 0.1096$, puis $D^+ = \max_{i=1,\ldots,n}\left(\frac{i}{n} - F_i\right) = 0.1176$, et enfin $D = \max\left(D^-, D^+\right) = 0.1176$;
6. nous comparons au seuil critique $D_{crit} = 0.161$ lue dans la table à 5%.

Dans notre exemple, $D < D_{crit}$, les données sont compatibles avec l'hypothèse de normalité.

Implémentations et logiciels

Les calculs étant relativement simples, ils ne diffèrent guère d'un logiciel à l'autre. Nous obtenons les mêmes résultats que sous un tableur.

Ce n'est guère étonnant. La seule différence pourrait résider dans le calcul de la loi de répartition $F(z)$. Mais il y a un consensus maintenant au niveau de l'implémentation de la loi normale. Les bibliothèques utilisées sont très similaires, voire identiques, les résultats sont forcément les mêmes.

12. Hervé Abdi et Paul Molin, *Lilliefors/Van Soet's test of normality*, In : Neil Salkind (Ed.) (2007), Encyclopedia of Measurement and Statistics ; accessible en ligne `http://www.utdallas.edu/~herve/Abdi-Lillie2007-pretty.pdf`
13. =LOI.NORMALE.STANDARD(...) sous EXCEL
14. `http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm`

	Moyenne	Ecart-type			
	3.4138	0.1968			
	x(i)	z(i)	F(Z)	D+	D-
1	3.109	-1.5484	0.0608	-0.0263	0.0608
2	3.114	-1.5230	0.0639	0.0051	0.0294
3	3.131	-1.4366	0.0754	0.0280	0.0064
4	3.140	-1.3909	0.0821	0.0558	-0.0213
5	3.178	-1.1978	0.1155	0.0569	-0.0224
6	3.186	-1.1572	0.1236	0.0833	-0.0488
7	3.203	-1.0708	0.1421	0.0993	-0.0648
8	3.223	-0.9692	0.1662	0.1096	-0.0752
9	3.288	-0.6390	0.2614	0.0489	-0.0144
10	3.378	-0.1817	0.4279	-0.0831	0.1176
11	3.391	-0.1156	0.4540	-0.0747	0.1091
12	3.395	-0.0953	0.4620	-0.0482	0.0827
13	3.414	0.0012	0.5005	-0.0522	0.0867
14	3.418	0.0215	0.5086	-0.0258	0.0603
15	3.440	0.1333	0.5530	-0.0358	0.0703
16	3.466	0.2654	0.6047	-0.0529	0.0874
17	3.481	0.3416	0.6337	-0.0475	0.0820
18	3.493	0.4026	0.6564	-0.0357	0.0702
19	3.493	0.4026	0.6564	-0.0012	0.0357
20	3.500	0.4382	0.6694	0.0203	0.0142
21	3.506	0.4687	0.6803	0.0438	-0.0093
22	3.506	0.4687	0.6803	0.0783	-0.0438
23	3.538	0.6312	0.7361	0.0570	-0.0226
24	3.558	0.7329	0.7682	0.0594	-0.0249
25	3.572	0.8040	0.7893	0.0728	-0.0383
26	3.605	0.9717	0.8344	0.0622	-0.0277
27	3.605	0.9717	0.8344	0.0966	-0.0622
28	3.770	1.8100	0.9649	0.0007	0.0338
29	3.898	2.4603	0.9931	0.0069	0.0275
			max	**0.1096**	**0.1176**

D	0.1176

Fig. 2.3. Test de Lilliefors avec un tableur

2.3 Test de Anderson-Darling

Description

Le test de Anderson-Darling est une autre variante du test de Kolmogorov-Smirnov, à la différence qu'elle donne plus d'importance aux queues de distribution [15]. De ce point de vue, elle est plus indiquée dans la phase d'évaluation des données précédant la mise en oeuvre d'un test paramétrique (comparaison de moyenne, de variances, etc.) que le test de Lilliefors.

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

Concernant l'adéquation à la loi normale, la statistique du test s'écrit :

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln(F_i) + \ln(1 - F_{n-i+1})\right] \tag{2.3}$$

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

Une correction est recommandée pour les petits effectifs [16], cette statistique corrigée est également utilisée pour calculer la *p-value* :

15. `http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm`
16. "Petits" étant assez vague, certains logiciels tel que STATISTICA ne valident l'utilisation du test d'Anderson-Darling uniquement pour $10 \leq n \leq 40$; la librairie intégrée dans R n'autorise pas le calcul lorsque $n < 8$

$$A_m = A\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \tag{2.4}$$

Les valeurs critiques A_{crit} pour différents niveaux de risques sont résumées dans le tableau suivant, ils ont été produits par simulation et ne dépendent pas de l'effectif de l'échantillon :

α	A_{crit}
0.10	0.631
0.05	0.752
0.01	1.035

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

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

Calculs

Moyenne 3.4138 Ecart-type 0.1968

	x(i)	z(i)	F(Z)	ln(F_i)	1-F_n-i+1	ln(1-F_n-i+1)	S
1	3.109	-1.5484	0.0608	-2.8008	0.0069	-4.9704	-7.7712
2	3.114	-1.5230	0.0639	-2.7508	0.0351	-3.3482	-18.2969
3	3.131	-1.4366	0.0754	-2.5848	0.1656	-1.7981	-21.9147
4	3.140	-1.3909	0.0821	-2.4995	0.1656	-1.7981	-30.0834
5	3.178	-1.1978	0.1155	-2.1586	0.2107	-1.5573	-33.4432
6	3.186	-1.1572	0.1236	-2.0907	0.2318	-1.4618	-39.0779
7	3.203	-1.0708	0.1421	-1.9511	0.2639	-1.3320	-42.6802
8	3.223	-0.9692	0.1662	-1.7944	0.3197	-1.1405	-44.0243
9	3.288	-0.6390	0.2614	-1.3416	0.3197	-1.1405	-42.1960
10	3.378	-0.1817	0.4279	-0.8488	0.3306	-1.1068	-37.1562
11	3.391	-0.1156	0.4540	-0.7897	0.3436	-1.0682	-39.0169
12	3.395	-0.0953	0.4620	-0.7721	0.3436	-1.0682	-42.3279
13	3.414	0.0012	0.5005	-0.6922	0.3663	-1.0043	-42.4111
14	3.418	0.0215	0.5086	-0.6761	0.3953	-0.9280	-43.3110
15	3.440	0.1333	0.5530	-0.5923	0.4470	-0.8053	-40.5306
16	3.466	0.2654	0.6047	-0.5031	0.4914	-0.7105	-37.6209
17	3.481	0.3416	0.6337	-0.4562	0.4995	-0.6941	-37.9606
18	3.493	0.4026	0.6564	-0.4210	0.5380	-0.6200	-36.4340
19	3.493	0.4026	0.6564	-0.4210	0.5460	-0.6051	-37.9656
20	3.500	0.4382	0.6694	-0.4014	0.5721	-0.5585	-37.4356
21	3.506	0.4687	0.6803	-0.3852	0.7386	-0.3030	-28.2158
22	3.506	0.4687	0.6803	-0.3852	0.8338	-0.1818	-24.3785
23	3.538	0.6312	0.7361	-0.3064	0.8579	-0.1533	-20.6882
24	3.558	0.7329	0.7682	-0.2637	0.8764	-0.1319	-18.5960
25	3.572	0.8040	0.7893	-0.2366	0.8845	-0.1227	-17.6071
26	3.605	0.9717	0.8344	-0.1811	0.9179	-0.0857	-13.6041
27	3.605	0.9717	0.8344	-0.1811	0.9246	-0.0784	-13.7513
28	3.770	1.8100	0.9649	-0.0358	0.9361	-0.0660	-5.5984
29	3.898	2.4603	0.9931	-0.0070	0.9392	-0.0627	-3.9700
						Somme	**-858.0675**

A	0.5885

A_m	0.6053
p-value	0.1159

Fig. 2.4. Test de Anderson-Darling avec un tableur

La mise en place du test passe par les étapes suivantes (Figure) :

1. les données sont triées pour former la série $x_{(i)}$;
2. nous estimons les paramètres, $\bar{x} = 3.4138$ et $s = 0.1968$;
3. nous calculons alors les données centrées et réduites $z_{(i)} = \frac{x_{(i)} - \bar{x}}{s}$;

4. nous utilisons la fonction de répartition de la normale centrée et réduite [17] pour obtenir les fréquences théoriques F_i ;
5. nous calculons la colonne $\ln(F_i)$;
6. de la même manière, nous formons F_{n-i+1} puis en déduisons $\ln(1 - F_{n-i+1})$;
7. nous calculons alors la somme $S = \sum_{i=1}^{n}(2i-1)\left[\ln(F_i) + \ln(1 - F_{n-i+1})\right] = -858.0675$;
8. la statistique $A = -n - \frac{1}{n}S = 0.5885$;
9. que nous comparons au seuil critique 0.752 à 5%.

Dans notre exemple, $A < A_{crit}$, l'hypothèse de normalité est compatible avec nos données.

Implémentations et logiciels

Logiciels

Les calculs étant relativement simples, ils ne varient guère d'un logiciel à l'autre. Dans notre exemple, TANAGRA et R fournissent la même valeur $A = 0.5885$. La différence est dans le calcul de la *p-value*. TANAGRA se contente de spécifier une plage de p-value en comparant la statistique aux seuils critiques relatifs aux différents niveaux de risque. Dans le cas présent, il indique p-value > 0.10 (Figure 2.1).

Calcul de la p-value

La p-value est calculée à partir de la statistique A_m par interpolation à partir d'une table décrite dans Stephens, M.A. (1986), *Tests based on EDF statistics*. In : D'Agostino, R.B. and Stephens, M.A., eds. : Goodness-of-Fit Techniques. Marcel Dekker, New York.

Nous donnons ici la règle de calcul implémentée dans le package **nortest** du logiciel R [18] :

1. calculer la statistique transformée $A_m = 0.6053$;
2. utiliser la règle suivante pour en déduire la *p-value*

A_m	p-value
$A_m < 0.2$	$1 - e^{-13.436+101.14\times A_m-223.73\times(A_m)^2}$
$0.2 \leq A_m < 0.34$	$1 - e^{-8.318+42.796\times A_m-59.938\times(A_m)^2}$
$0.34 \leq A_m < 0.6$	$e^{0.9177-4.279\times A_m-1.38\times(A_m)^2}$
$0.66 \leq A_m$	$e^{1.2937-5.709\times A_m+0.0186\times(A_m)^2}$

3. nous obtenons ainsi p-value $= 0.1159$, cohérent avec le résultat indiqué par TANAGRA.

17. =LOI.NORMALE.STANDARD(...) sous EXCEL
18. `http://www.biostat.wustl.edu/archives/html/s-news/2005-04/msg00065.html`

2.4 Test de D'Agostino

Description

Le test de D'Agostino [19], connu également sous l'appellation *test $K2$ (K-squared) de D'Agostino-Pearson*, est basé sur les coefficients d'asymétrie et d'aplatissement. Lorsque ces deux indicateurs diffèrent *simultanément* de la valeur de référence 0, on conclut que la distribution empirique n'est pas compatible avec la loi normale. L'enjeu est de construire une combinaison efficace de ces indicateurs.

L'idée est très simple à comprendre, sa puissance est considérée comme très bonne au point que son auteur préconise de le substituer aux tests basés sur la statistique de Kolmogorov-Smirnov. Le test de D'Agostino présenterait une puissance similaire à celle de Shapiro-Wilk à mesure que les effectifs augmentent. Il devient particulièrement efficace à partir de $n \geq 20$, on le préfère alors aux tests basés sur la statistique de Kolmogorov-Smirnov [20]. Par rapport au test de Shapiro-Wilk, il serait de surcroît peu sensible à l'existence des ex-aequo dans l'échantillon.

Le reproche usuellement adressé au test de D'Agostino est qu'il ne permet pas directement de comprendre la nature de la déviation de la loi normale en cas de rejet de l'hypothèse nulle. Il faut compléter l'analyse avec l'étude individuelle des coefficients, ou en mettant en oeuvre les techniques descriptives décrites précédemment.

Si l'idée est simple, les formules sont relativement complexes. Il faut procéder par étapes. Le fil directeur est de centrer et réduire les deux coefficients (asymétrie et aplatissement) de manière à obtenir des valeurs z_1 et z_2 distribuées asymptotiquement selon une loi normale $\mathcal{N}(0,1)$. La transformation intègre des corrections supplémentaires de manière à rendre l'approximation normale plus efficace.

Transformation du coefficient d'asymétrie

Une première transformation est effectuée sur le coefficient d'asymétrie. Les calculs successifs sont les suivants [21].

19. `http://en.wikipedia.org/wiki/D'Agostino's_K-squared_test`, j'ai néanmoins quelques doutes sur les formules fournies en ligne.

20. Voir Zar J.H. (1996) - *Biostatistical Analysis* - Prentice Hall International Editions. ; une description est disponible en ligne `http://calamar.univ-ag.fr/uag/staps/cours/stat/stat.htm`

21. Voir implémentation MATLAB tirée de Trujillo-Ortiz, A. and R. Hernandez-Walls. (2003). DagosPtest : D'Agostino-Pearson's K2 test for assessing normality of data using skewness and kurtosis. A MATLAB file., `http://www.mathworks.com/matlabcentral/files/3954/DagosPtest.m`

$$
\begin{aligned}
g_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}} \\
A &= g_1\sqrt{\frac{(n+1)(n+3)}{6(n-2)}} \\
B &= \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} \\
C &= \sqrt{2(B-1)}-1 \\
D &= \sqrt{C} \\
E &= \frac{1}{\sqrt{\ln(D)}} \\
F &= \frac{A}{\sqrt{\frac{2}{C-1}}} \\
z_1 &= E\ln\left(F+\sqrt{F^2+1}\right)
\end{aligned}
$$

Transformation du coefficient d'aplatissement

Nous procédons de manière similaire pour le coefficient d'aplatissement.

$$
\begin{aligned}
g_2 = G_2 &= \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_i\left(\frac{x_i-\bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \\
G &= \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \\
H &= \frac{(n-2)(n-3)g_2}{(n+1)(n-1)\sqrt{G}} \\
J &= \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}} \\
K &= 6+\frac{8}{J}\left[\frac{2}{J}+\sqrt{1+\frac{4}{J^2}}\right] \\
L &= \left(\frac{1-\frac{2}{K}}{1+H\sqrt{\frac{2}{K-4}}}\right) \\
z_2 &= \frac{(1-\frac{2}{9K})-L^{\frac{1}{3}}}{\sqrt{\frac{2}{9K}}}
\end{aligned}
$$

z_1 et z_2 suivent tous deux asymptotiquement une loi normale $\mathcal{N}(0,1)$. La statistique du test est la combinaison

$$K2 = z_1^2 + z_2^2 \tag{2.5}$$

Elle suit asymptotiquement une loi du χ^2 à 2 degrés de libertés. L'incompatibilité de la distribution évaluée avec la loi normale est d'autant plus marquée que la statistique $K2$ prend une valeur élevée. Pour un risque α, la région critique du test s'écrit :

$$R.C. : K2 > \chi^2_{1-\alpha}(2)$$

Pour $\alpha = 0.05$, le seuil critique est $\chi^2_{0.95}(2) = 5.99$.

Calculs

Pour notre ensemble de données, les calculs s'articulent comme suit (Figure 2.5) :

Moyenne 3.4138

	x(i)	d=x-x_bar	(x-x_bar)^2	(x-x_bar)^3	(x-x_bar)^4
1	3.109	-0.3048	0.0929	-0.0283	0.0086
2	3.114	-0.2998	0.0899	-0.0269	0.0081
3	3.131	-0.2828	0.0800	-0.0226	0.0064
4	3.140	-0.2738	0.0749	-0.0205	0.0056
5	3.178	-0.2358	0.0556	-0.0131	0.0031
6	3.186	-0.2278	0.0519	-0.0118	0.0027
7	3.203	-0.2108	0.0444	-0.0094	0.0020
8	3.223	-0.1908	0.0364	-0.0069	0.0013
9	3.288	-0.1258	0.0158	-0.0020	0.0003
10	3.378	-0.0358	0.0013	0.0000	0.0000
11	3.391	-0.0228	0.0005	0.0000	0.0000
12	3.395	-0.0188	0.0004	0.0000	0.0000
13	3.414	0.0002	0.0000	0.0000	0.0000
14	3.418	0.0042	0.0000	0.0000	0.0000
15	3.440	0.0262	0.0007	0.0000	0.0000
16	3.466	0.0522	0.0027	0.0001	0.0000
17	3.481	0.0672	0.0045	0.0003	0.0000
18	3.493	0.0792	0.0063	0.0005	0.0000
19	3.493	0.0792	0.0063	0.0005	0.0000
20	3.500	0.0862	0.0074	0.0006	0.0001
21	3.506	0.0922	0.0085	0.0008	0.0001
22	3.506	0.0922	0.0085	0.0008	0.0001
23	3.538	0.1242	0.0154	0.0019	0.0002
24	3.558	0.1442	0.0208	0.0030	0.0004
25	3.572	0.1582	0.0250	0.0040	0.0006
26	3.605	0.1912	0.0366	0.0070	0.0013
27	3.605	0.1912	0.0366	0.0070	0.0013
28	3.770	0.3562	0.1269	0.0452	0.0161
29	3.898	0.4842	0.2345	0.1135	0.0550

g1	0.208142
A	0.506686
B	3.563811
C	1.264426
D	1.124467
E	2.919674
F	0.184236
Z1	0.534913

g2	-0.005306
G	0.498971
H	-0.006277
J	1.733577
K	18.369529
L	0.89321578
Z2	0.225892

K2	0.3372
p-value	0.8449

Fig. 2.5. Test de D'Agostino avec un tableur

1. calculer la moyenne empirique $\bar{x} = 3.4148$;
2. former la colonne $d = x - \bar{x}$;
3. puis les colonnes d^2, d^3 et d^4 ;
4. calculer successivement les valeurs énumérées ci-dessus pour aboutir à $z_1 = 0.5349$ et $z_2 = 0.2259$;
5. nous formons la statistique $K2 = 0.5349^2 + 0.2259^2 = 0.3372$;
6. nous pouvons également calculer la p-value à l'aide de la fonction de répartition [22] du χ^2, et obtenir p-value $= 0.8449$.

Dans notre exemple, la statistique $K2$ est largement inférieure à $\chi^2_{0.95}(2) = 5.99$, la distribution observée est compatible avec une distribution théorique normale.

22. =LOI.KHIDEUX(...) dans EXCEL

Implémentations et logiciels

La procédure se résume à des calculs arithmétiques simples que l'on peut facilement implémenter. En revanche la complexité des formules incite à la prudence, des coquilles peuvent facilement s'immiscer. On peut se poser des questions quant à la fiabilité des sources proposées en ligne.

Nous avons testé les données sur TANAGRA et R [package **fBasic**, fonction dagoTest(...)]. Nous avons obtenu des résultats concordants. Même si ça n'a pas valeur de preuve, c'est quand même un signe positif. Nous retrouvons exactement les valeurs calculées dans le tableur.

2.5 Test de Jarque-Bera

Description

Le test de normalité de Jarque-Bera [23] est également fondé sur les coefficients d'asymétrie et d'aplatissement. Il évalue les écarts simultanés de ces coefficients avec les valeurs de référence de la loi normale. La formulation est très simple par rapport au test de D'Agostino, le prix est une puissance moindre. Il ne devient réellement intéressant que lorsque les effectifs sont élevés.

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

On propose les estimateurs

$$b_1 = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_i (x_i - \bar{x})^2\right)^{\frac{3}{2}}} \tag{2.6}$$

$$b_2 = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_i (x_i - \bar{x})^2\right)^{2}} \tag{2.7}$$

La loi conjointe de ces estimateurs est normale bivariée, on écrit [24] :

$$\sqrt{n}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix}\right] \tag{2.8}$$

La matrice de variance covariance présentée ici est une expression simplifiée valable pour les grandes valeurs de n. Il est possible de produire des expressions plus précises, affichées par les logiciels de statistique. Nous notons également que la covariance de b_1 et b_2 est nulle.

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

23. http://en.wikipedia.org/wiki/Jarque-Bera_test
24. http://wis.kuleuven.be/stat/robust/Papers/tailweightCOMPSTAT04.pdf

$$T = n\left(\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right) \tag{2.9}$$

Elle est distribuée asymptotiquement selon une loi du χ^2 à 2 degrés de liberté, tout comme la statistique de D'Agostino vue précédemment.

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

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

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

En vérité, ce test est toujours moins puissant que le test de D'Agostino c.-à-d. il a une propension plus élevée à conclure à la compatibilité avec la loi normale. On devrait donc toujours préférer ce dernier. Dans la pratique, les écarts de puissance s'amenuisent à mesure que les effectifs augmentent. La simplicité des calculs, très faciles à appréhender et à mettre en oeuvre sur des outils simples tels qu'un tableur, militent en faveur du test de Jarque-Bera.

Calculs

Pour notre ensemble de données, les calculs s'articulent comme suit (Figure 2.6) :

Moyenne
3.4138

	x(i)	x-x_bar	(x-x_bar)^2	(x-x_bar)^3	(x-x_bar)^4
1	3.109	-0.3048	0.0929	-0.0283	0.0086
2	3.114	-0.2998	0.0899	-0.0269	0.0081
3	3.131	-0.2828	0.0800	-0.0226	0.0064
4	3.140	-0.2738	0.0749	-0.0205	0.0056
5	3.178	-0.2358	0.0556	-0.0131	0.0031
6	3.186	-0.2278	0.0519	-0.0118	0.0027
7	3.203	-0.2108	0.0444	-0.0094	0.0020
8	3.223	-0.1908	0.0364	-0.0069	0.0013
9	3.288	-0.1258	0.0158	-0.0020	0.0003
10	3.378	-0.0358	0.0013	0.0000	0.0000
11	3.391	-0.0228	0.0005	0.0000	0.0000
12	3.395	-0.0188	0.0004	0.0000	0.0000
13	3.414	0.0002	0.0000	0.0000	0.0000
14	3.418	0.0042	0.0000	0.0000	0.0000
15	3.440	0.0262	0.0007	0.0000	0.0000
16	3.466	0.0522	0.0027	0.0001	0.0000
17	3.481	0.0672	0.0045	0.0003	0.0000
18	3.493	0.0792	0.0063	0.0005	0.0000
19	3.493	0.0792	0.0063	0.0005	0.0000
20	3.500	0.0862	0.0074	0.0006	0.0001
21	3.506	0.0922	0.0085	0.0008	0.0001
22	3.506	0.0922	0.0085	0.0008	0.0001
23	3.538	0.1242	0.0154	0.0019	0.0002
24	3.558	0.1442	0.0208	0.0030	0.0004
25	3.572	0.1582	0.0250	0.0040	0.0006
26	3.605	0.1912	0.0366	0.0070	0.0013
27	3.605	0.1912	0.0366	0.0070	0.0013
28	3.770	0.3562	0.1269	0.0452	0.0161
29	3.898	0.4842	0.2345	0.1135	0.0550

b1	0.2081
b2	2.7956
jB	0.2599
p-value	0.8781

Fig. 2.6. Test de Jarque-Bera avec un tableur

1. calculer la moyenne empirique $\bar{x} = 3.4148$;
2. former la colonne $d = x - \bar{x}$;
3. puis les colonnes d^2, d^3 et d^4 ;

4. calculer successivement les estimateurs $b_1 = 0.2081$ et $b_2 = 2.7956$;
5. nous formons la statistique $T = 0.2081^2 + 2.7956^2 = 0.2599$;
6. nous pouvons également calculer la p-value à l'aide de la fonction de répartition [25] du χ^2, et obtenir p-value $= 0.8781$.

Dans notre exemple, la statistique T est largement inférieure à $\chi^2_{0.95}(2) = 5.99$, la distribution observée est compatible avec une distribution théorique normale.

Nous observons que la p-value du test est très similaire à celle fournie par le test de D'Agostino.

Implémentations et logiciels

La simplicité des calculs facilite sa diffusion. Nous avons testé la fonction *jarqueBeraTest(...)* du package **fBasic** dans R. Nous avons obtenu exactement des résultats identiques à ceux du tableur.

2.6 Conclusion sur les tests de normalité

Dans ce document nous avons présenté les techniques destinées à évaluer la compatibilité d'une distribution empirique avec la loi normale.

Sans mésestimer la pertinence des tests statistiques, on pouvait en pressentir les résultats à la lumière des graphiques de distribution (Histogramme de fréquences, figure 1.2), de la boîte à moustaches (Figure 1.3) et de la droite de Henry (Figure 1.7). Nous avions de plus une idée sur la nature des désaccords : la distribution est unimodale, très légèrement asymétrique, et la droite de Henry nous indique que les écart sont essentiellement situées dans les queues de distribution.

Il est quand même heureux que les caractéristiques des tests étudiées ici confirment cette impression. Certes, ils aboutissent tous à la même conclusion, la compatibilité avec la loi normale, mais à des degrés différents que l'on peut appréhender à l'aide de la p-value. On constate que le test le moins enclin à accepter l'hypothèse de normalité est celui de Anderson-Darling, qui est justement sensible aux écarts dans les queues de distribution (p-value $= 0.1159$).

Le test de Shapiro-Wilk propose une p-value égale 0.1408. Sa puissance est reconnue dans la littérature, elle confirme cette idée ici.

Le test de Lilliefors, pourtant fondé sur la statistique de Kolmogorov-Smirnov, est plus conservateur, avec un p-value de 0.3859. Cela est surtout consécutif à la nature du désaccord, en queue de distribution, qu'elle détecte mal.

Enfin, les tests de D'Agostino et de Jarque-Bera, basés sur les coefficients d'asymétrie et d'aplatissement acceptent volontiers l'hypothèse de normalité avec une p-value > 0.8. Elles souffrent vraisemblablement de la petite taille de notre fichier $(n = 29)$ [26].

25. =LOI.KHIDEUX(...) dans EXCEL
26. Voir l'article de Sneyers (1974) pour une stratégie de choix des tests de normalité.

Finalement, pourquoi étudier à l'aide de tests compliqués ce que l'on pouvait appréhender sur des graphiques simples ? Les tests amènent un point de vue objectif, avec une approche rigoureuse. C'est un argument fort lorsque nos résultats font l'objet d'enjeux importants. De plus, lorsque nous avons à traiter un grand nombre de variables, il est intéressant de disposer d'outils automatisés pour tester un grand nombre de variables, quitte à revenir attentivement sur les variables qui posent problème par la suite. Mais pour cela, il faut comprendre le comportement des outils que l'on utilise.

Autre aspect très important, la détection et le traitement des points atypiques que nous avons introduits au début de ce document n'était pas du tout anodin dans notre contexte. La vérification de la normalité passe par l'estimation de l'espérance mathématique, paramètre de la loi normale, à l'aide de la moyenne empirique. La présence de points douteux peuvent fausser totalement les calculs, et par conséquent les conclusions du test.

La suppression de ces points comme nous l'avons réalisé est une solution possible. Adopter des estimations robustes de l'espérance en est une autre.

3

Tests de symétrie

Dans certains cas, on peut se contenter de tester la symétrie d'une distribution. Le test est bien entendu moins restrictif puisqu'il ne porte que sur un aspect de la forme de la distribution.

Exemple 1. Pour évaluer une régression multiple de la forme $Y = f(X;\theta)+\epsilon$, une distribution des résidus $\hat{\epsilon}$ asymétrique laisse à penser que le modèle est mal spécifié. Le graphique des résidus est un outil important, nous pouvons également mettre en oeuvre des tests statistiques.

Exemple 2. Dans certaines techniques non-paramétriques, le test de Wilcoxon pour échantillons appariés [1] par exemple, la symétrie est requise pour que le test agisse correctement.

3.1 Test de symétrie basé sur le coefficient d'asymétrie

Description

Un test de symétrie fondé sur le coefficient d'asymétrie est la première stratégie qui vient à l'esprit. Il s'agit d'utiliser une partie du test de D'Agostino ou de Jarque-Bera.

La statistique du test asymptotique que nous proposons utilise la première composante du test de Jarque-Bera :

$$b_1 = g_1 = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_i (x_i - \bar{x})^2\right)^{\frac{3}{2}}} \tag{3.1}$$

Pour une meilleure efficacité, nous produisons une estimation de la variance plus précise

$$\sigma_1^2 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \tag{3.2}$$

Sous l'hypothèse nulle de distribution normale, le rapport $\frac{b_1}{\sigma_1}$ suit asymptotiquement une loi $\mathcal{N}(0,1)$.

La région critique du test s'écrit :

1. http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

$$R.C. : |\frac{b_1}{\sigma_1}| > u_{1-\frac{\alpha}{2}}$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ lue dans la table de la loi normale centrée-réduite.

Il s'agit d'une distribution asymptotique. Mais ce test peut être utilisé pour des effectifs relativement faibles. On le conseille[2] généralement pour $8 \leq n \leq 5000$.

Calculs

Dans notre classeur EXCEL, les calculs s'articulent de la manière suivante (Figure 3.1) :

1. calculer la moyenne empirique $\bar{x} = 3.4148$;
2. former la colonne $d = x - \bar{x}$;
3. puis les colonnes d^2 et d^3 ;
4. calculer successivement $b_1 = 0.2081$, $\sigma_1^2 = 0.1880$ et $\sigma_1 = 0.4335$;
5. nous formons la statistique $|\frac{b_1}{\sigma_1}| = 0.4801$;
6. que nous comparons au seuil critique $u_{0.975} = 1.96$

Moyenne
3.4138

	x(i)	x-x_bar	(x-x_bar)^2	(x-x_bar)^3
1	3.109	-0.3048	0.0929	-0.0283
2	3.114	-0.2998	0.0899	-0.0269
3	3.131	-0.2828	0.0800	-0.0226
4	3.140	-0.2738	0.0749	-0.0205
5	3.178	-0.2358	0.0556	-0.0131
6	3.186	-0.2278	0.0519	-0.0118
7	3.203	-0.2108	0.0444	-0.0094
8	3.223	-0.1908	0.0364	-0.0069
9	3.288	-0.1258	0.0158	-0.0020
10	3.378	-0.0358	0.0013	0.0000
11	3.391	-0.0228	0.0005	0.0000
12	3.395	-0.0188	0.0004	0.0000
13	3.414	0.0002	0.0000	0.0000
14	3.418	0.0042	0.0000	0.0000
15	3.440	0.0262	0.0007	0.0000
16	3.466	0.0522	0.0027	0.0001
17	3.481	0.0672	0.0045	0.0003
18	3.493	0.0792	0.0063	0.0005
19	3.493	0.0792	0.0063	0.0005
20	3.500	0.0862	0.0074	0.0006
21	3.506	0.0922	0.0085	0.0008
22	3.506	0.0922	0.0085	0.0008
23	3.538	0.1242	0.0154	0.0019
24	3.558	0.1442	0.0208	0.0030
25	3.572	0.1582	0.0250	0.0040
26	3.605	0.1912	0.0366	0.0070
27	3.605	0.1912	0.0366	0.0070
28	3.770	0.3562	0.1269	0.0452
29	3.898	0.4842	0.2345	0.1135

b1	0.2081
v(b1)	0.1880
sigma1	0.4335

statistique	0.4801
p-value	0.6312

G1	0.2197

Fig. 3.1. Test de symétrie basé sur le coefficient d'asymétrie

Les données sont compatibles avec une symétrie gaussienne. Bien entendu, disposant de la loi de répartition de la statistique du test, nous pouvons calculer la p-value, elle est égale à 0.6312.

2. Voir Tassi, *Méthodes Statistiques*, Economica, 1992, pages 323-324.

Implémentations et logiciels

A l'instar de l'estimation de la variance sur un échantillon, plutôt que l'estimation triviale du coefficient d'asymétrie, qui est biaisée, les logiciels produisent une estimation non-biaisée[3].

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^3$$
$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

L'estimation de l'écart type de la statistique n'est pas modifiée, nous utilisons toujours σ_1.

Dans notre exemple, $G_1 = 0.2197$, le rapport $\frac{G_1}{\sigma_1} = 0.5068$. Ce sont les valeurs fournies par les logiciels STATISTICA, SPSS et TANAGRA.

Nous constatons que la conclusion du test n'est pas modifiée concernant notre exemple.

3.2 Test de symétrie - Test de Wilcoxon

Description

Le test de symétrie ci-dessus introduit une restriction qui peut être rédhibitoire : l'hypothèse nulle correspond à une distribution normale. Or, il se peut que l'on veuille couvrir une palette de distributions plus large. L'hypothèse que l'on veut tester est la compatibilité avec une loi symétrique, que ce soit une loi normale bien évidemment, mais aussi une loi de Laplace (exponentielle bilatérale), une loi de Cauchy, une loi uniforme, etc. Cette hypothèse de symétrie peut être importante dans certaines procédures statistiques.

Il nous faut donc définir une nouvelle statistique dont la distribution ne repose pas sur la loi de X. Nous nous tournons bien évidemment vers les tests non paramétriques.

Le test de symétrie que nous présentons dans cette section (Aïvazian, pages 322 à 325)[4], outre l'abandon de l'hypothèse de normalité, est plus général que le test précédent dans le sens où il permet d'examiner la symétrie par rapport à un point quelconque θ. L'hypothèse nulle de symétrie par rapport à θ s'écrit :

$$H_0 : f(x+\theta) = f(x-\theta) \tag{3.3}$$

où $f(\cdot)$ est la fonction de densité de la variable aléatoire étudiée.

Le test est fondé sur les rangs absolus par rapport à la valeur de référence θ. Concrètement, la procédure consiste à :

– calculer la variable $z_i = |x_i - \theta|$;

3. http ://en.wikipedia.org/wiki/Skewness
4. Wilcoxon F., *Individual comparaisons by ranking methods*, Biometrics, 1, 80-83, 1945.

- former la série triée $z_{(i)}$;
- pour chaque individu i tel que $x_i - \theta > 0$ (notons I^+ l'ensemble des individus répondant à cette condition), obtenir dans la série $z_{(i)}$ son rang r_i ;

La statistique du test est définie par

$$S^+ = \sum_{i \in I^+} r_i \tag{3.4}$$

Sous l'hypothèse nulle, nous pouvons obtenir l'espérance et la variance de S^+ :

$$E(S^+) = \frac{1}{4} n(n+1) \tag{3.5}$$

$$V(S^+) = \frac{1}{24} n(n+1)(2n+1) \tag{3.6}$$

Remarque 7 (Un cas particulier du test de comparaison de populations). Ce test peut se comprendre comme le test de Wilcoxon de comparaison de populations sur échantillons appariés. L'un des échantillons est formé par la valeur constante θ.

La statistique S^+ a été tabulée sous H_0 pour les petites valeurs de n. Lorsque les effectifs augmentent ($n \geq 15$ dans la pratique [5]), la quantité $U_s = \frac{S^+ - E(S^+)}{\sqrt{V(S^+)}}$ suit asymptotiquement une loi $\mathcal{N}(0,1)$. La région critique du test s'écrit :

$$R.C. : U_s = \frac{|S^+ - E(S^+)|}{\sqrt{V(S^+)}} > u_{1-\frac{\alpha}{2}}$$

où $u_{1-\frac{\alpha}{2}}$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la table de la loi normale centrée réduite.

Il est possible de calculer la *p-value* du test à partir de la fonction de répartition de la loi normale.

Remarque 8 (Correction de continuité). Pour une meilleure approximation, nous pouvons introduire la correction de continuité, la région critique s'écrit dans le cas :

$$R.C : U_s^- = \frac{S^+ - E(S^+) - 0.5}{\sqrt{V(S^+)}} < u_{\frac{\alpha}{2}} \text{ ou } U_s^+ = \frac{S^+ - E(S^+) + 0.5}{\sqrt{V(S^+)}} > u_{1-\frac{\alpha}{2}}$$

La région critique est réduite, le test est plus conservateur. La correction devient néanmoins négligeable à mesure que les effectifs augmentent.

Moyenne
3.4138

	x(i)	z_i=ABS(x-x_bar)	w_i=(x_i > x_bar)	Rang r_i	S+
1	3.109	0.3048	0	27	0
2	3.114	0.2998	0	26	0
3	3.131	0.2828	0	25	0
4	3.140	0.2738	0	24	0
5	3.178	0.2358	0	23	0
6	3.186	0.2278	0	22	0
7	3.203	0.2108	0	21	0
8	3.223	0.1908	0	18	0
9	3.288	0.1258	0	15	0
10	3.378	0.0358	0	6	0
11	3.391	0.0228	0	4	0
12	3.395	0.0188	0	3	0
13	3.414	0.0002	1	1	1
14	3.418	0.0042	1	2	2
15	3.440	0.0262	1	5	5
16	3.466	0.0522	1	7	7
17	3.481	0.0672	1	8	8
18	3.493	0.0792	1	9	9
19	3.493	0.0792	1	9	9
20	3.500	0.0862	1	11	11
21	3.506	0.0922	1	12	12
22	3.506	0.0922	1	12	12
23	3.538	0.1242	1	14	14
24	3.558	0.1442	1	16	16
25	3.572	0.1582	1	17	17
26	3.605	0.1912	1	19	19
27	3.605	0.1912	1	19	19
28	3.770	0.3562	1	28	28
29	3.898	0.4842	1	29	29
				Somme	**218**

E(S+)	217.5000
V(S+)	2138.7500

U_s	0.0108
p-value	0.9914

U-	0.0000
U+	0.0216
p-value	0.9914

Fig. 3.2. Test de symétrie de Wilcoxon

Calculs

Dans notre classeur EXCEL, les calculs s'articulent de la manière suivante (Figure 3.2) :

1. nous prenons comme valeur de référence la moyenne empirique $\theta = \bar{x} = 3.4138$;
2. former la série $z_i = |x_i - \bar{x}|$;
3. détecter les observations i pour lesquelles $x_i > \bar{x}$, nous avons créé une colonne de variable indicatrice w_i pour cela ;
4. calculer le rang r_i de chaque observation dans z_i ;
5. en déduire alors la statistique $S^+ = \sum_i w_i \times r_i = 218$, somme des rangs des observations pour lesquelles $x_i > \bar{x}$;
6. former la statistique centrée et réduite $U_s = 0.0108$;
7. que l'on compare au fractile $u_{0.975} = 1.96$ de la loi normale centrée et réduite.

Au risque de 5%, l'hypothèse de symétrie de la distribution des données est acceptée. Nous pouvons également calculer la p-value, elle est égale à 0.9914.

Les résultats ne sont guère affectés par la correction de continuité (Figure 3.2), la p-value est identique (jusqu'à la 4-ème décimale).

Traitement des ex aequo

Lorsque deux ou plusieurs observations présentent la même valeur, nous devons définir une stratégie pour l'affectation des rangs.

5. http ://www.chups.jussieu.fr/polys/biostats/poly/POLY.Chp.12.2.html

Valeur	1.2	2.4	2.4	2.4	3.7	3.7
Rang	1	3	2	4	6	5

Tableau 3.1. Traitement des ex aequo - Méthode des rangs aléatoires

Méthode des rangs aléatoires

La méthode des rangs aléatoires consiste, pour une valeur repérée plusieurs fois dans le fichier, à affecter aléatoirement un rang pris parmi les rangs attribués à la valeur. Dans notre exemple (Tableau 3.1) comportant 6 observations, 3 observations présentent la même valeur 2.4. Les rangs $\{2, 3, 4\}$ doivent être distribués aléatoirement à ces observations. L'intérêt de cette approche est que tout le processus décrit ci-dessus reste valable, notamment la formule de la variance. Son inconvénient est qu'il est moins puissant que la technique que nous décrirons plus bas. Autre reproche que l'on pourrait lui faire, l'exécution des calculs avec des générateurs de nombres aléatoires différents (par exemple lorsque le générateur est indexé sur l'horloge de la machine) peut aboutir à des conclusions contradictoires. Ce qui ne manque pas de plonger le non initié dans un abîme de perplexité.

Méthode des rangs moyens

La méthode des rangs moyens a le mérite de toujours fournir la même réponse. Elle est surtout plus puissante. Il s'agit, pour des observations portant la même valeur, de leur affecter un rang moyen. Dans notre exemple (Tableau 3.2), nous affectons $\frac{2+3+4}{3} = 3.0$ aux individus correspondant à la valeur 2.4, et $\frac{5+6}{2} = 5.5$ pour les individus correspondant à 3.7. La statistique, sa loi de distribution asymptotique, et son espérance ne sont pas modifiés. En revanche, il faut adapter la formule de la variance qui devient [6] :

$$V'(S^+) = \frac{n(n+1)(2n+1)}{24} - \frac{1}{2}\sum_{j=1}^{g} t_j(t_j-1)(t_j+1) \tag{3.7}$$

où g est le nombre de valeurs différentes dans le fichier, t_j est le nombre d'observations correspondant à une valeur. Dans notre exemple (Tableau 3.2), $g = 3$ pour les valeurs $\{1.2, 2.4, 3.7\}$, et les t_j sont $\{1, 3, 2\}$.

Remarque 9. Si $g = n$ (et donc $t_j = 1, \forall j$), il n'y a pas d'ex aequo dans le fichier de données, nous observons que les variances coïncident $V(S^+) = V'(S^+)$.

Valeur	1.2	2.4	2.4	2.4	3.7	3.7
Rang	1	3.0	3.0	3.0	5.5	5.5

Tableau 3.2. Traitement des ex aequo - Méthode des rangs moyens

6. Siegel S., Castellan Jr., J., *Nonparametric Statistics for Behavioral Science*, McGraw-Hill, 1988, page 94.

3.3 Test de symétrie - Test de Van der Waerden

Description

Ce test est une variante plus puissante du test de Wilcoxon, il introduit une légère modification de la statistique qui s'écrit maintenant

$$V^+ = \sum_{i \in I^+} \psi\left(\frac{1}{2} + \frac{1}{2}\frac{r_i}{n+1}\right) \tag{3.8}$$

$\psi(\cdot)$ est la fonction inverse de la loi normale centrée réduite.

L'espérance et la variance de V^+ s'écrivent

$$E(V^+) = \frac{1}{2}\sum_{k=1}^{n} \psi\left(\frac{1}{2} + \frac{1}{2}\frac{k}{n+1}\right) \tag{3.9}$$

$$V(V^+) = \frac{1}{4}\sum_{k=1}^{n} \left[\psi\left(\frac{1}{2} + \frac{1}{2}\frac{k}{n+1}\right)\right]^2 \tag{3.10}$$

Sous H_0, la quantité $U_v = \frac{V^+ - E(V^+)}{\sqrt{V(V^+)}}$ suit asymptotiquement une loi normale $\mathcal{N}(0,1)$. La définition de la région critique est similaire à celle du test de Wilcoxon.

Calculs

Moyenne 3.4138

	x(i)	z_i=ABS(x-x_bar)	w_i=(x_i > x_bar)	Rang r_i	a_i	b	V+	c	d	e
1	3.109	0.3048	0	27	0.9500	1.6449	0.0000	0.5167	0.0418	0.0017
2	3.114	0.2998	0	26	0.9333	1.5011	0.0000	0.5333	0.0837	0.0070
3	3.131	0.2828	0	25	0.9167	1.3830	0.0000	0.5500	0.1257	0.0158
4	3.140	0.2738	0	24	0.9000	1.2816	0.0000	0.5667	0.1679	0.0282
5	3.178	0.2358	0	23	0.8833	1.1918	0.0000	0.5833	0.2104	0.0443
6	3.186	0.2278	0	22	0.8667	1.1108	0.0000	0.6000	0.2533	0.0642
7	3.203	0.2108	0	21	0.8500	1.0364	0.0000	0.6167	0.2967	0.0881
8	3.223	0.1908	0	18	0.8000	0.8416	0.0000	0.6333	0.3407	0.1161
9	3.288	0.1258	0	15	0.7500	0.6745	0.0000	0.6500	0.3853	0.1485
10	3.378	0.0358	0	6	0.6000	0.2533	0.0000	0.6667	0.4307	0.1855
11	3.391	0.0228	0	4	0.5667	0.1679	0.0000	0.6833	0.4770	0.2276
12	3.395	0.0188	0	3	0.5500	0.1257	0.0000	0.7000	0.5244	0.2750
13	3.414	0.0002	1	1	0.5167	0.0418	0.0418	0.7167	0.5730	0.3283
14	3.418	0.0042	1	2	0.5333	0.0837	0.0837	0.7333	0.6229	0.3880
15	3.440	0.0262	1	5	0.5833	0.2104	0.2104	0.7500	0.6745	0.4549
16	3.466	0.0522	1	7	0.6167	0.2967	0.2967	0.7667	0.7279	0.5299
17	3.481	0.0672	1	8	0.6333	0.3407	0.3407	0.7833	0.7835	0.6139
18	3.493	0.0792	1	9	0.6500	0.3853	0.3853	0.8000	0.8416	0.7083
19	3.493	0.0792	1	9	0.6500	0.3853	0.3853	0.8167	0.9027	0.8149
20	3.500	0.0862	1	11	0.6833	0.4770	0.4770	0.8333	0.9674	0.9359
21	3.506	0.0922	1	12	0.7000	0.5244	0.5244	0.8500	1.0364	1.0742
22	3.506	0.0922	1	12	0.7000	0.5244	0.5244	0.8667	1.1108	1.2338
23	3.538	0.1242	1	14	0.7333	0.6229	0.6229	0.8833	1.1918	1.4204
24	3.558	0.1442	1	16	0.7667	0.7279	0.7279	0.9000	1.2816	1.6424
25	3.572	0.1582	1	17	0.7833	0.7835	0.7835	0.9167	1.3830	1.9127
26	3.605	0.1912	1	19	0.8167	0.9027	0.9027	0.9333	1.5011	2.2533
27	3.605	0.1912	1	19	0.8167	0.9027	0.9027	0.9500	1.6449	2.7055
28	3.770	0.3562	1	28	0.9667	1.8339	1.8339	0.9667	1.8339	3.3632
29	3.898	0.4842	1	29	0.9833	2.1280	2.1280	0.9833	2.1280	4.5286
						Somme	**11.1716**		**22.5427**	**26.1101**

E(V+)	11.2714
V(V+)	6.5275
U_v	0.0391
p-value	0.9688

Fig. 3.3. Test de symétrie de Van der Waerden

Dans notre classeur EXCEL, les calculs s'articulent de la manière suivante (Figure 3.3) :

1. nous prenons comme valeur de référence la moyenne empirique $\theta = \bar{x} = 3.4138$;
2. former la série $z_i = |x_i - \bar{x}|$;
3. détecter les observations i pour lesquelles $x_i > \bar{x}$, nous avons créé une colonne de variable indicatrice w_i pour cela ;
4. calculer le rang r_i de chaque observation dans z_i ;
5. produire la valeur $a_i = \left(\frac{1}{2} + \frac{1}{2}\frac{r_i}{n+1}\right)$;
6. puis, $b_i = \psi(a_i)$;
7. en déduire alors la statistique $V^+ = \sum_i w_i \times b_i = 11.1716$, somme des rangs transformés des observations pour lesquelles $x_i > \bar{x}$;
8. pour calculer l'espérance mathématique et la variance de V^+, nous formons les colonnes $c_i = \left(\frac{1}{2} + \frac{1}{2}\frac{k}{n+1}\right)$, $d_i = \psi(c_i)$ et $e_i = d_i^2$;
9. nous calculons $E(V^+) = \frac{1}{2}\sum_i d_i = 11.2714$ et $V(V^+) = \frac{1}{4}\sum_i e_i = 6.5275$;
10. reste à produire $U_v = \frac{|11.1716 - 11.2714|}{\sqrt{6.5275}} = 0.0391$;
11. que l'on compare au fractile $u_{0.975} = 1.96$ de la loi normale centrée et réduite.

Au risque de 5%, l'hypothèse de symétrie de la distribution des données est acceptée. Nous pouvons également calculer la p-value, elle est égale à 0.9688.

Traitement des ex aequo

Comme précédemment, en adoptant la méthode des rangs moyens, nous devons adapter la formule de la variance :

$$V'(V^+) = \frac{1}{4}\sum_{j=1}^{g} t_j \left[\psi\left(\frac{1}{2} + \frac{1}{2}\frac{\bar{r}_j}{n+1}\right)\right]^2 \tag{3.11}$$

où $\bar{r}_j$ est le rang moyen pour le j-ème groupe de valeur.

Remarque 10. Encore une fois, s'il n'y a pas d'ex aequo (c.-à-d. $t_j = 1, \forall j$), les variances coïncident $V(V^+) = V'(V^+)$.

3.4 Conclusion sur les tests de symétrie

Lorsque le paramètre θ n'est pas fourni par la théorie (par ex. la moyenne des résidus est égale à 0 dans la régression linéaire multiple avec constante) ou la connaissance du domaine, il nous faut l'estimer. Différents paramètres de localisation peuvent être utilisées, la moyenne arithmétique comme nous avons pu le faire dans ce support, mais aussi des paramètres moins sensibles aux données atypiques telles que la médiane. Quoiqu'il en soit, lorsque θ est estimé, les procédures non paramétriques présentées dans ce chapitre sont approximatifs (Aïvazian, page 324).

Enfin, bien que leur utilité soit indéniable, les tests de symétrie sont curieusement absents de la plupart des logiciels de statistique.

4

Transformation de Box-Cox

Une grande partie des procédures statistiques reposent sur la normalité des distributions. Et quand bien même certains d'entre eux seraient assez robustes, on sait généralement que des distributions très dissymétriques faussent les calculs, notamment les techniques basées sur des distances entre individus, ou pire des distances par rapport à la moyenne. Transformer les variables de manière à se rapprocher de la distribution normale, ou tout du moins pour les symétriser, est parfois un préalable nécessaire avant toute analyse statistique.

Il est possible de rendre gaussienne toute variable aléatoire continue par une transformation monotone continue. Les fonctions les plus répandues sont certainement $y = \sqrt{x}$ et $y = \ln(x)$ ([1], page 275). Mais le résultat laisse parfois à désirer, poussant les utilisateurs à empiler au petit bonheur la chance les transformations. Il faut adopter une démarche raisonnée.

4.1 Fonctions de transformation de Box-Cox

Box et Cox proposent des fonctions de transformations plus génériques, car paramétrables. En les modulant au mieux, nous pouvons nous rapprocher de la distribution normale. Deux types de fonctions sont généralement décrites dans la littérature [1].

La première propose un seul paramètre λ :

$$y = \phi(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \ln(x) & (\lambda = 0) \end{cases} \tag{4.1}$$

La seconde, plus générale, mais plus difficile à appréhender, propose 2 paramètres λ_1 et λ_2 :

$$y = \varphi(x) = \begin{cases} \frac{(x+\lambda_2)^{\lambda_1} - 1}{\lambda_1} & (\lambda_1 \neq 0) \\ \ln(x + \lambda_2) & (\lambda_1 = 0) \end{cases} \tag{4.2}$$

La principale difficulté est de préciser la bonne valeur des paramètres sur un échantillon de données. C'est en cela que la première formulation est plus accessible (équation 4.1), nous ne manipulons qu'un seul paramètre. Dans ce qui suit, nous montrons comment, en pratique, nous fixons de manière efficace la valeur adéquate du paramètre λ.

1. `http://en.wikipedia.org/wiki/Box-Cox_transformation`

4.2 Approche graphique : utiliser la Droite de Henry

4.2.1 La droite de Henry

Le Q-Q plot (quantile-quantile plot) consiste à produire un graphique "nuage de points" où : nous plaçons en abcisse les quantiles observés, en ordonnée les quantiles théoriques de la loi normale (section 1.5, nous laissons directement les quantiles de loi normale centrée réduite en ordonnée dans ce chapitre). Si les points forment une droite, la distribution empirique est compatible avec la loi normale.

Prenons un exemple pour fixer les idées. Nous analysons le montant des crédits (X) accordés par une banque à un échantillon de $n = 50$ clients. Nous cherchons à savoir si la distribution est normale. Le plus simple est de produire le graphique Q-Q plot (Figure 4.1) en suivant les prescriptions décrites par ailleurs (section 1.5). Nous observons dans la feuille de calcul :

	A	B	C	D
1				
2	n	50		
3				
4	X	i	F	U
5	385	1	0.01	-2.24
6	672	2	0.03	-1.85
7	727	3	0.05	-1.62
8	804	4	0.07	-1.46
9	976	5	0.09	-1.33
10	1028	6	0.11	-1.22
11	1037	7	0.13	-1.12
12	1038	8	0.15	-1.03
13	1131	9	0.17	-0.95
14	1155	10	0.19	-0.87
15	1199	11	0.21	-0.80
16	1288	12	0.23	-0.73
17	1299	13	0.25	-0.67
18	1412	14	0.27	-0.61
19	1455	15	0.29	-0.55
20	1480	16	0.31	-0.49
21	1495	17	0.33	-0.44
22	1647	18	0.35	-0.38
23	1835	19	0.37	-0.33
24	1893	20	0.39	-0.28
25	1935	21	0.41	-0.23
26	1984	22	0.43	-0.18
27	2116	23	0.45	-0.13
28	2186	24	0.47	-0.07
29	2221	25	0.49	-0.02
30	2247	26	0.51	0.02
31	2273	27	0.53	0.07
32	2320	28	0.55	0.13
33	2326	29	0.57	0.18
34	2353	30	0.59	0.23
35	2359	31	0.61	0.28
36	2427	32	0.63	0.33
37	2528	33	0.65	0.38
38	2631	34	0.67	0.44
39	3104	35	0.69	0.49
40	3108	36	0.71	0.55
41	3160	37	0.73	0.61
42	3181	38	0.75	0.67
43	3386	39	0.77	0.73
44	3394	40	0.79	0.80
45	3422	41	0.81	0.87
46	3651	42	0.83	0.95
47	3832	43	0.85	1.03
48	3913	44	0.87	1.12
49	5152	45	0.89	1.22
50	5293	46	0.91	1.33
51	5842	47	0.93	1.46
52	6289	48	0.95	1.62
53	7596	49	0.97	1.85
54	7758	50	0.99	2.24

r
0.9272

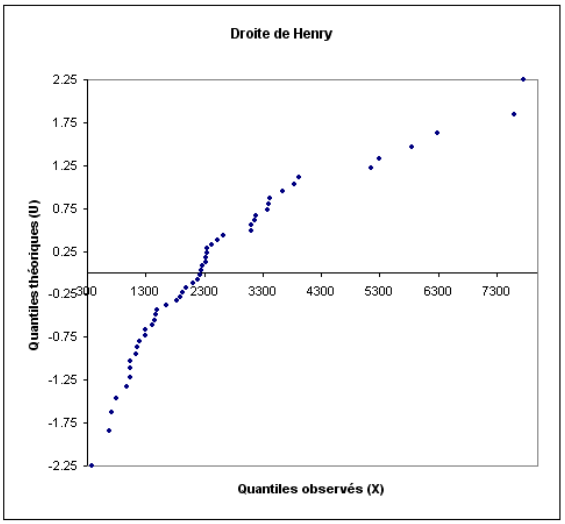


Fig. 4.1. Droite de Henry - Données non transformées

– Dans la colonne **A**, nous avons les valeurs de X, triées de manière croissante. Ces valeurs correspondent donc aux quantiles.

- La colonne **B** sert uniquement à numéroter les observations. En **C**, nous avons la fonction de répartition empirique selon la formule $F_i = \frac{i-0.375}{n+0.25}$
- Nous utilisons l'inverse de la loi normale centrée réduite[2] pour produire la série u_i
- Le graphique Q-Q plot est formé par les couples (x_i, u_i). On constate dans notre cas que les points ne sont pas alignés sur une droite. L'hypothèse de normalité n'est pas crédible. Il nous faut transformer la variable X pour nous rapprocher de la distribution normale.

4.2.2 Exploiter la droite de Henry

Une idée très simple est de créer une nouvelle colonne Y basée sur la transformation de Box-Cox (équation 4.1) dans la feuille de calcul, en réservant une cellule à part pour le paramètre λ. Nous reconstruisons le graphique quantile-quantile sur Y. Nous pouvons alors tâtonner en fixant différentes valeurs de λ, nous observons à chaque fois la situation de la droite de Henry. Nous arrêtons les itérations lorsque nous obtenons un résultat satisfaisant. Séduisante a priori, cette démarche est très vite fastidieuse, elle ne nous permet pas de tester un grand nombre de valeurs de λ. La situation devient intenable si nous avons un grand nombre de variables à traiter.

Comment exploiter au mieux les informations fournies par la droite de Henry tout en ayant la possibilité de tester un nombre élevé de valeurs ?

Pour répondre à cette question, il nous faut proposer un critère numérique qui donne des indications sur le caractère linéaire de la série de points du graphique Q-Q plot. Il en existe un, qui n'est pas fait pour ça, mais qui répond très bien à la spécification : le coefficient de corrélation linéaire de Pearson[3]. En effet, le coefficient de corrélation r indique l'intensité de la liaison linéaire entre deux variables. Plus les points seront alignés dans le graphique quantile-quantile, plus la valeur de r se rapprochera de $+1$. Dans le cas idéal, les points forment une droite, nous aurons $r = +1$.

Bien évidemment, nous dévoyons un peu le coefficient de corrélation. La valeur de r n'a pas sens dans notre contexte. Il ne faut pas chercher à l'interpréter. Nous ne souhaitons pas mesurer l'association entre les quantiles théoriques et les quantiles observés. Il s'agit avant tout d'un critère destiné à caractériser l'alignement des points. Dans notre exemple ci-dessus (Figure 4.1), nous avons $r = 0.9272$. Est-ce qu'il est possible de produire une variable Y, en utilisant la formule 4.1, de manière à augmenter encore cette valeur ?

4.2.3 Box-Cox Normality Plot

La bonne stratégie pour détecter facilement la valeur adéquate du paramètre dans la transformation est donc de balayer un grand nombre de valeurs de λ, et de surveiller la valeur de r calculée sur la droite de Henry. On choisira la valeur λ^* qui maximise r.

2. LOI.NORMALE.STANDARD.INVERSE(.) de EXCEL
3. `http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf`

Pour obtenir une vue synthétique de la simulation, on construit généralement un graphique[4] qui met en relation λ (en abcisse) et r (en ordonnée), il s'agit du *Box-Cox Normality Plot*[5]

Application sur notre exemple (Figure 4.1) Nous avons reconstruit notre feuille EXCEL en rajoutant la colonne Y paramétrée par λ. Nous insérons dans une des cellules la formule du coefficient de corrélation linéaire calculé sur les couples de points (y_i, u_i). A l'aide de l'outil "Table de simulation" d'EXCEL, nous avons produit les séries de valeurs (λ, r) pour $\lambda = -2$ à $+2$ avec un pas de 0.1. Nous reproduisons le tableau des valeurs et le graphique *Box-Cox Normality Plot* (Figure 4.2).

Lambda	
-2	0.65744
-1.9	0.67860
-1.8	0.70045
-1.7	0.72286
-1.6	0.74567
-1.5	0.76872
-1.4	0.79179
-1.3	0.81465
-1.2	0.83707
-1.1	0.85879
-1	0.87956
-0.9	0.89914
-0.8	0.91731
-0.7	0.93386
-0.6	0.94861
-0.5	0.96143
-0.4	0.97221
-0.3	0.98089
-0.2	0.98744
-0.1	0.99186
0	0.99420
0.1	0.99451
0.2	0.99290
0.3	0.98948
0.4	0.98437
0.5	0.97774
0.6	0.96973
0.7	0.96051
0.8	0.95023
0.9	0.93908
1	0.92720
1.1	0.91474
1.2	0.90186
1.3	0.88868
1.4	0.87533
1.5	0.86190
1.6	0.84850
1.7	0.83520
1.8	0.82208
1.9	0.80918
2	0.79657

Max	0.99451

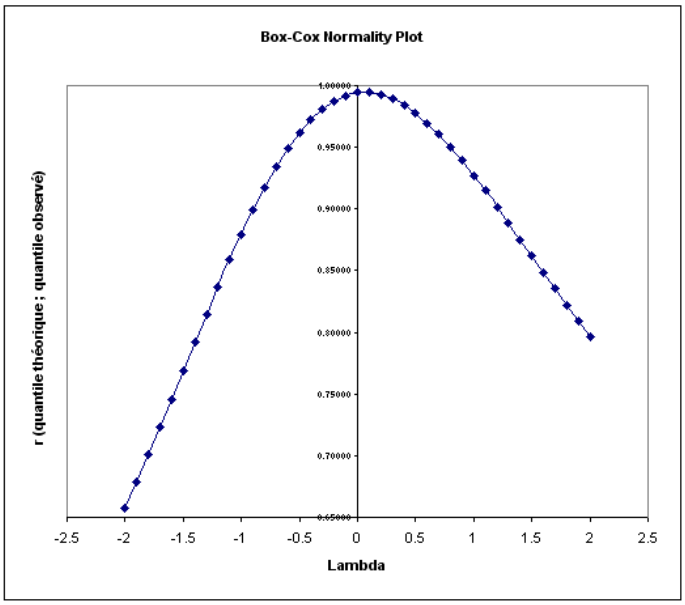


Fig. 4.2. *Box-Cox Normality Plot*

Dans notre exemple, il semble que la bonne valeur soit $\lambda^* = 0.1$. Elle maximise la corrélation entre quantile théoriques et quantiles observés dans la droite de Henry avec $r^* = 0.99451$. Essayons de reporter cela dans notre feuille de calcul pour visualiser le graphique quantile-quantile de la variable transformée

4. Voir `http://www.itl.nist.gov/div898/handbook/eda/section3/boxcoxno.htm`

5. N.A. : Malgré mes recherches, je n'ai pas réussi à trouver l'équivalent francophone. Je ne voulais pas me lancer dans un néologisme sorti de nulle part. Si un lecteur statisticien connaît l'appellation appropriée en français, j'accueillerai avec beaucoup de plaisir ses indications.

$$y = \frac{x^{0.1} - 1}{0.1}$$

Voici le détail de la nouvelle feuille de calcul (Figure 4.3) :

	A	B	C	D	E
1					
2	n	50		Lambda	0.1
3					
4	X	i	F	U	Y
5	385	1	0.01	-2.24	8.14
6	672	2	0.03	-1.85	9.18
7	727	3	0.05	-1.62	9.33
8	804	4	0.07	-1.46	9.52
9	976	5	0.09	-1.33	9.90
10	1028	6	0.11	-1.22	10.01
11	1037	7	0.13	-1.12	10.03
12	1038	8	0.15	-1.03	10.03
13	1131	9	0.17	-0.95	10.20
14	1155	10	0.19	-0.87	10.24
15	1199	11	0.21	-0.80	10.32
16	1288	12	0.23	-0.73	10.46
17	1299	13	0.25	-0.67	10.48
18	1412	14	0.27	-0.61	10.65
19	1455	15	0.29	-0.55	10.72
20	1480	16	0.31	-0.49	10.75
21	1495	17	0.33	-0.44	10.77
22	1647	18	0.35	-0.38	10.97
23	1835	19	0.37	-0.33	11.20
24	1893	20	0.39	-0.28	11.27
25	1935	21	0.41	-0.23	11.31
26	1984	22	0.43	-0.18	11.37
27	2116	23	0.45	-0.13	11.51
28	2186	24	0.47	-0.07	11.58
29	2221	25	0.49	-0.02	11.61
30	2247	26	0.51	0.02	11.64
31	2273	27	0.53	0.07	11.66
32	2320	28	0.55	0.13	11.70
33	2326	29	0.57	0.18	11.71
34	2353	30	0.59	0.23	11.74
35	2359	31	0.61	0.28	11.74
36	2427	32	0.63	0.33	11.80
37	2528	33	0.65	0.38	11.89
38	2631	34	0.67	0.44	11.98
39	3104	35	0.69	0.49	12.35
40	3108	36	0.71	0.55	12.35
41	3160	37	0.73	0.61	12.39
42	3181	38	0.75	0.67	12.40
43	3386	39	0.77	0.73	12.54
44	3394	40	0.79	0.80	12.55
45	3422	41	0.81	0.87	12.56
46	3651	42	0.83	0.95	12.71
47	3832	43	0.85	1.03	12.82
48	3913	44	0.87	1.12	12.87
49	5152	45	0.89	1.22	13.51
50	5293	46	0.91	1.33	13.57
51	5842	47	0.93	1.46	13.80
52	6289	48	0.95	1.62	13.98
53	7596	49	0.97	1.85	14.44
54	7758	50	0.99	2.24	14.49

r
0.9945

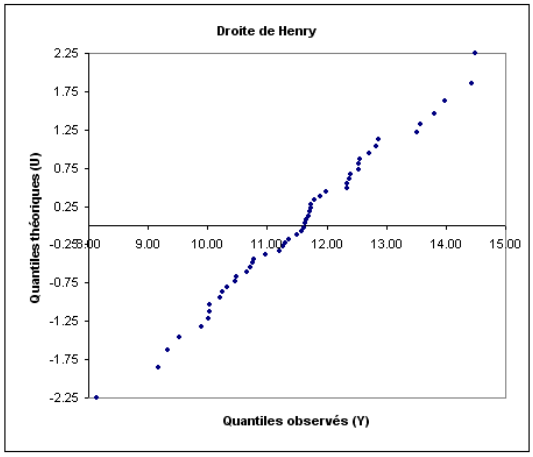


Fig. 4.3. Droite de Henry - Données transformées

- Les colonnes **A** à **D** sont identiques à la feuille initiale (Figure 4.1).
- Dans la colonne **E**, nous insérons la variable Y, paramétrée par λ en cellule **E2**.
- A la lumière des valeurs testées précédemment, nous fixons $\lambda = 0.1$
- Nous construisons alors le graphique Q-Q plot à l'aide des quantiles observées (y_i, en abcisse) et des quantiles théoriques (u_i, en ordonnée)
- Par rapport aux données initiales, les points sont maintenant mieux alignés, assimilable à une droite. La distribution de Y se rapproche de la loi normale.

4.2.4 Tester la normalité

Pour valider notre démarche, utilisons les tests de normalité mis en avant dans le chapitre 2. Nous souhaitons vérifier l'efficacité de la transformation en testant la compatibilité de X, puis de Y, avec la

distribution normale. Nous verrons ainsi si l'analyse graphique guidée par la Droite de Henry produit des résultats cohérents avec les procédures statistiques.

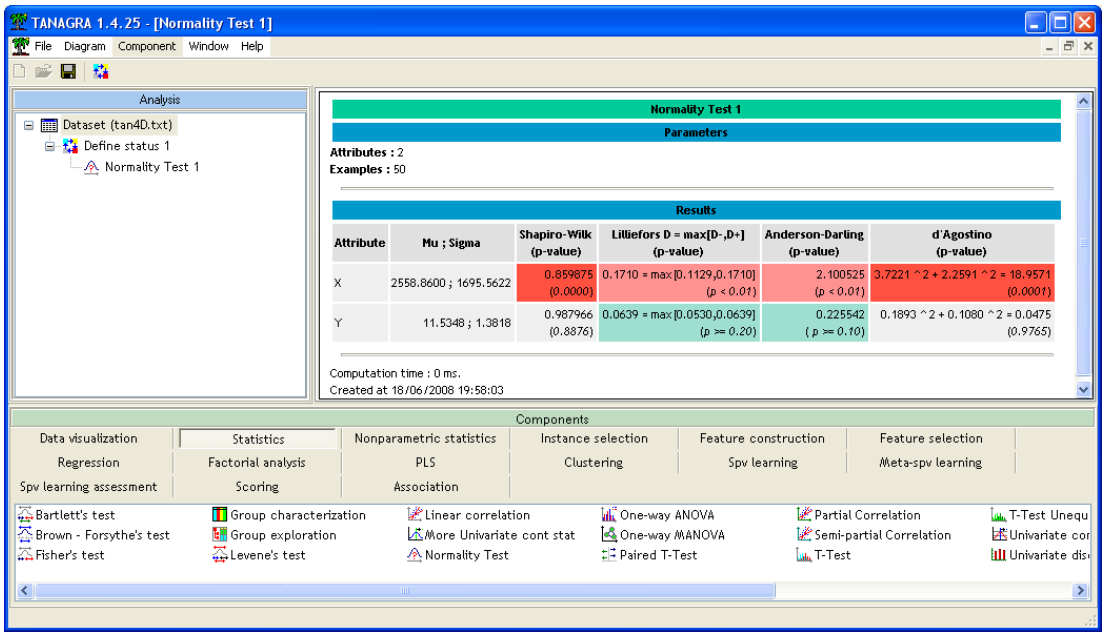


Fig. 4.4. Tests de normalité - Données avant et après transformation

Nous utilisons le logiciel TANAGRA [6]. Au niveau de signification 5%, quel que soit le test utilisé, il apparaît clairement que X n'est pas gaussienne. Après la transformation de Box-Cox (équation 4.1) avec $\lambda = 0.1$, la variable modifiée Y est compatible avec la loi normale (Figure 4.4). Manifestement, l'opération a été réalisée avec succès.

Malgré tout, il reste un bémol. La stratégie mise en oeuvre repose sur le tâtonnement. La plage de valeurs à tester doit être définie judicieusement. Le risque de passer à côté de la valeur optimale n'est pas négligeable. De plus, l'analyse graphique se prête mal à un traitement d'un grand nombre de variables. Pour ces raisons, nous présentons dans la section suivante une démarche fondée sur un processus d'optimisation que l'on peut automatiser.

4.3 Approche numérique : la maximisation de la vraisemblance

4.3.1 Fonction de densité des variables Y et X

La variable transformée Y est distribuée normalement. Elle est paramétrée par λ si on s'en tient à la première formulation (Equation 4.1). On peut s'appuyer sur le principe du maximum de vraisemblance

6. Voir `http://tutoriels-data-mining.blogspot.com/2008/04/tests-dadquation-la-loi-normale.html` concernant la mise en oeuvre de ces tests dans le logiciel

pour produire la valeur optimale λ^*. Pour un échantillon de taille n, nous souhaitons maximiser la vraisemblance :

$$L = \prod_{i=1}^{n} f(y_i)$$

où $f(.)$ est la fonction de densité de la loi normale.

En pratique, pour des raisons de commodités numériques, on procède plutôt à l'optimisation de la log-vraisemblance

$$LL = \sum_{i=1}^{n} \ln f(y_i)$$

Y est elle-même exprimée à partir de X, il faudrait revenir à la fonction de densité $g(x)$. Il existe un lien entre les fonctions de densités lorsque une des variable est fonction d'une autre. La formule générique est la suivante [7] :

$$g(x) = f(y) \times |\frac{\partial y}{\partial x}| \tag{4.3}$$

où $\frac{\partial y}{\partial x} = \phi'(x)$ est la dérivée partielle première par rapport à X de la fonction $\phi(x)$.

Dans le cas de l'équation 4.1, son logarithme s'exprime de manière très simple :

$$\ln \frac{\partial y}{\partial x} = \ln \phi'(x) = (\lambda - 1) \ln x \tag{4.4}$$

4.3.2 Expression de la fonction à optimiser / λ

En prenant en compte toutes ces informations, nous pouvons écrire la fonction de log-vraisemblance que nous devons optimiser par rapport à λ. Nous utilisons directement les estimateurs usuels en ce qui concerne les autres paramètres (moyenne et écart-type).

La log-vraisemblance s'écrit

$$LL = -\frac{n}{2} \ln(2\pi) - n \times \ln(s) - \frac{1}{2s^2} \sum_i [\phi(x_i) - m]^2 + (\lambda - 1) \sum_i \ln x_i \tag{4.5}$$

avec

$$m = \frac{1}{n} \sum_i \phi(x_i)$$
$$s^2 = \frac{1}{n-1} \sum_i [\phi(x_i) - m]^2$$

La valeur λ^* qui maximise la log-vraisemblance produit la transformation souhaitée. Y se rapproche au mieux de la distribution normale.

7. Voir par exemple `http://rfv.insa-lyon.fr/~jolion/STAT/node32.html`. Attention, notre situation est inversée, nous connaissons la distribution de la variable transformée, nous souhaitons revenir à la fonction de densité de la variable initiale

Remarque 11 (Simplification de LL). Tout ce qui ne dépend pas de λ peut être retirée de l'expression 4.5. On peut ainsi omettre le premier terme de la formule.

Remarque 12 (Intervalle de confiance de λ). Il est possible de produire un intervalle de variation de λ pour un niveau de confiance $(1-\alpha)$. Le calcul est fondé sur le principe du rapport de vraisemblance. L'idée est de définir la plage de valeurs de λ où la variable transformée Y est compatible avec la loi normale [8].

4.3.3 Application numérique

Nous allons essayer de tirer profit du SOLVEUR d'EXCEL pour obtenir la valeur optimale du paramètre λ. Mais auparavant, il nous préparer la feuille de calcul, notamment en produisant la log-vraisemblance. La cellule de cette dernière doit bien entendu être dépendante de la cellule de λ.

Détaillons la feuille EXCEL (Figure) :

- En colonne **A**, nous avons les données originelles x.
- En colonne **B**, nous avons les données transformées $y = \phi(x)$, paramétré par λ en **B2**
- En **B56** et **B57**, nous avons respectivement m et s
- A partir de **D5**, nous produisons les valeurs individuelles de la log-vraisemblance c.-à-d.

$$l(x_i) = -\frac{1}{2s^2}[y_i - m]^2 + (\lambda - 1)\ln x_i$$

- Il ne nous reste plus qu'à réaliser la somme pour obtenir la log-vraisemblance en **D2**,

$$LL = -\frac{n}{2}\ln(2\pi) - n\ln(s) + \sum_i l(x_i)$$

- Pour $\lambda = 1$, nous obtenons $LL = -442.235$
- En lançant l'outil SOLVEUR, **D2** en cellule cible et **B2** en cellule variable, nous obtenons

$$\lambda^{**} = 0.06603$$

avec

$$LL^{**} = -430.878$$

Nous ne manquons pas de comparer cette valeur avec celle obtenue par tâtonnement, nous avions trouvé $\lambda^* = 0.1$. Mais bien entendu le résultat était tributaire de la précision que nous avions définie lors du processus de recherche. Avec l'approche par maximisation de la vraisemblance, le résultat est obtenu directement, elle peut être automatisée. Cette caractéristique est particulièrement intéressante dès lors que nous avons à traiter un grand nombre de variables.

8. Voir `http://www.itl.nist.gov/div898/handbook/pmc/section5/pmc52.htm`. A ce sujet, je me suis rendu compte que l'expression de la log-vraisemblance sur ce site semble erronée. C'est assez étrange. De manière générale, NIST fait référence. Quoiqu'il en soit, sur l'exemple proposé en ligne (*Example of Box-Cox scheme*), en utilisant la feuille de calcul que nous présenterons dans la section suivante, basée sur la formule 4.5, nous retrouvons le bon résultat $\lambda^* = 0.276$ avec $LL = 46.918$. En utilisant leur formulation, le processus d'optimisation ne converge pas.

	A	B	C	D
1	n	Lambda		LL
2	50	0.06603		-430.878
3				
4	X	Y		LL_i
5	385	7.29		-8.731
6	672	8.13		-7.575
7	727	8.25		-7.459
8	804	8.41		-7.327
9	976	8.71		-7.130
10	1028	8.80		-7.090
11	1037	8.81		-7.084
12	1038	8.81		-7.084
13	1131	8.95		-7.032
14	1155	8.98		-7.022
15	1199	9.04		-7.006
16	1288	9.16		-6.985
17	1299	9.17		-6.983
18	1412	9.30		-6.973
19	1455	9.35		-6.973
20	1480	9.38		-6.975
21	1495	9.40		-6.976
22	1647	9.55		-6.996
23	1835	9.73		-7.045
24	1893	9.78		-7.064
25	1935	9.82		-7.079
26	1984	9.86		-7.098
27	2116	9.96		-7.152
28	2186	10.02		-7.183
29	2221	10.05		-7.199
30	2247	10.06		-7.211
31	2273	10.08		-7.224
32	2320	10.12		-7.247
33	2326	10.12		-7.250
34	2353	10.14		-7.263
35	2359	10.15		-7.266
36	2427	10.19		-7.301
37	2528	10.26		-7.354
38	2631	10.33		-7.410
39	3104	10.61		-7.686
40	3108	10.61		-7.688
41	3160	10.64		-7.720
42	3181	10.65		-7.733
43	3386	10.76		-7.859
44	3394	10.76		-7.864
45	3422	10.77		-7.882
46	3651	10.89		-8.026
47	3832	10.97		-8.141
48	3913	11.01		-8.193
49	5152	11.48		-8.986
50	5293	11.53		-9.075
51	5842	11.71		-9.420
52	6289	11.84		-9.696
53	7596	12.18		-10.478
54	7758	12.21		-10.572
55				
56	m	9.98		
57	s	1.07		

Fig. 4.5. Transformation de Box-Cox - Maximisation de la vraisemblance

Toujours au titre de la comparaison des résultats, nous avons calculé la corrélation entre les quantiles théoriques et les quantiles observés de la droite de Henry avec la transformation $\lambda^{**} = 0.06603$, nous obtenons $r^{**} = 0.99462$, à comparer avec la valeur $r^{*} = 0.9945$ obtenue lors de la recherche par tâtonnement. Le gain est relativement faible quand même. L'avantage comparatif ne se situe pas à ce niveau là.

Remarque 13 (Optimisation numérique). Obtenir une expression directe de λ en fonction des x_i pour l'optimisation de la log-vraisemblance serait l'idéal. Mais ce n'est pas possible. Nous sommes obligés de nous tourner vers des procédures d'optimisation numérique. La fonction SOLVEUR en est l'illustration parfaite. Elle est capable de produire une solution en lui fournissant uniquement la fonction objectif et les paramètres à manipuler. Pour les connaisseurs, le SOLVEUR est basé, dixit la documentation Microsoft,

sur l'algorithme GRG (*Generalized Reduced Gradient*) [9]. Pour ma part, j'ai rarement vu un outil aussi souple et fiable.

Remarque 14 (Calcul des paramètres λ_1 et λ_2 pour la fonction de transformation $\varphi(x)$). Concernant la fonction $\varphi(x)$, nous pouvons adopter les deux démarches ci-dessus :

1. Tâtonner en fournissant des plages de valeurs de λ_1 et λ_2 à tester. Le graphique *Normal Probability Plot* résumant les différentes configurations sera en 3D, mais ça n'est en rien rédhibitoire.
2. Optimiser la fonction de vraisemblance en fonction de λ_1 et λ_2. Le processus est exactement le même, l'information à connaître pour écrire convenablement la log-vraisemblance est

$$\ln \frac{\partial y}{\partial x} = (\lambda_1 - 1) \ln(x + \lambda_2)$$

9. Voir `http://support.microsoft.com/kb/214115/en-us/`

A

Gestion des versions

Un support de cours n'est jamais figé, nous essayons constamment de les enrichir. Dans cette annexe, nous recensons les différentes versions de ce document.

Version 1.0 Première version mise en ligne, au mois d'Août 2007. Il comprend les chapitres 1, 2 et 3.

Version 2.0 Le chapitre 4 a été intégré au document en Juin 2008.

B

Mise en oeuvre des tests de normalité dans TANAGRA

TANAGRA est un logiciel open source accessible en ligne. Il implémente plusieurs techniques d'exploration de données issues de la statistique, de la statistique exploratoire et de la fouille de données (Data Mining). Au-delà du logiciel, une série de didacticiels sont disponibles, accompagnées de jeu de données. L'idée est de présenter brièvement les enjeux de la méthode, proposer un jeu de données test, et montrer la démarche à suivre avec le logiciel.

Les tests d'adéquation à la loi normale sont implémentés dans le composant NORMALITY TEST situé dans l'onglet STATISTICS. Plusieurs tests sont disponibles : le test de Shapiro-Wilk, de Lilliefors, d'Anderson-Darling et de D'Agostino (Figure B.1).

Listons quelques références utiles concernant le test de normalité :

- `http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html`, URL du site ;
- `http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Normality_Test.pdf`, didacticiel détaillant la mise en oeuvre du test de normalité sur un jeu de données ;
- `http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/normality_test_simulation.xls`, données utilisées pour le didacticiel.

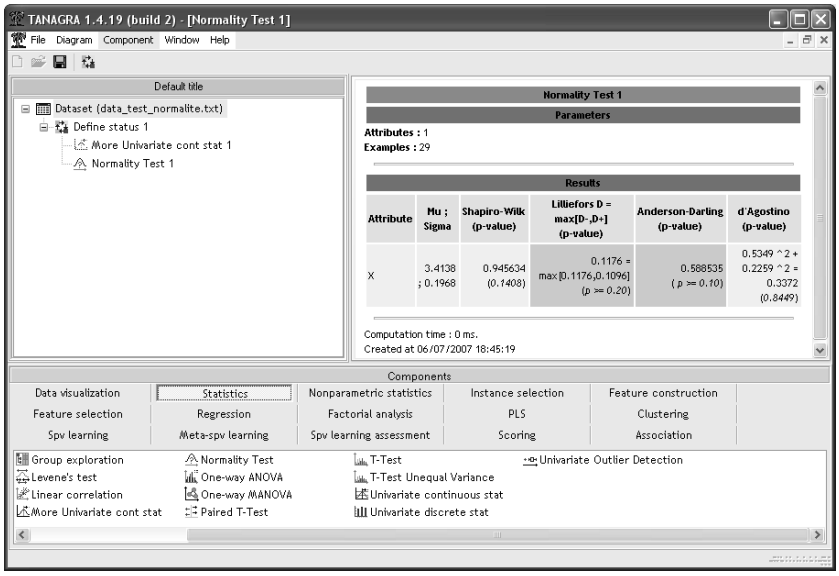


Fig. B.1. Copie d'écran du logiciel TANAGRA

C

Code source et packages R pour les tests de normalité

Le logiciel R (`http://www.r-project.org/`) est un interpréteur de commandes doté d'un vrai langage de programmation, et possédant une bibliothèque très riche de techniques statistiques.

Tout un chacun peut programmer une fonction correspondant à telle ou telle nouvelle technique. De plus, le principe des packages est un autre dispositif qui permet d'enrichir considérablement le logiciel. Les utilisateurs peuvent produire des bibliothèques externes spécialisées que l'on peut intégrer facilement.

S'agissant des tests d'adéquation à la loi normale, nous avons téléchargé et installé les packages **nortest** et **fBasics**. Tous les tests décrits dans ce support ont pu être évalués sur notre jeu de données.

Listing C.1. Code source pour R

```
#vider la mémoire de tous les objets
rm(list = ls())
#modifier le répertoire de travail et charger les données
setwd("...")
data <- read.csv(file="data_test_normalite.csv")
#copier les données dans un vecteur
x <- data$X
#le test de Shapiro-Wilk
shapiro.test(x)
#*** charger le package "nortest" ***
library(nortest)
#test de Lilliefors
lillie.test(x)
#test d'Anderson-Darling
ad.test(x)
#*** charger la librairie "fBasics"
library(fBasics)
#test de D'Agostino
dagoTest(x)
#test de Jarque-Bera
jarqueberaTest(x)
```

Listing C.2. Sorties de R

```
> shapiro.test(x) #le test de Shapiro-Wilk
        Shapiro-Wilk normality test

data:  x
W = 0.9456, p-value = 0.1408

> library(nortest)
> lillie.test(x) #test de Lilliefors
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  x
D = 0.1176, p-value = 0.3859

> ad.test(x) #test d'Anderson-Darling
        Anderson-Darling normality test

data:  x
A = 0.5885, p-value = 0.1159

> library(fBasics)
> dagoTest(x) #test de D'Agostino
Title:
 D'Agostino Normality Test

Test Results:
  STATISTIC:
    Chi2 | Omnibus: 0.3372
    Z3   | Skewness: 0.5349
    Z4   | Kurtosis: 0.2259
  P VALUE:
    Omnibus  Test: 0.8449
    Skewness Test: 0.5927
    Kurtosis Test: 0.8213

> jarqueberaTest(x) #test de Jarque-Bera
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 0.2599
  P VALUE:
    Asymptotic p Value: 0.8781
```

Littérature

1. Aïvazian, S., Enukov, I., Mechalkine, L., *Eléments de modélisation et traitement primaire des données*, Mir, 1986.
2. Borcard, D., *Tests de normalité*, `http://biol10.biol.umontreal.ca/BIO2042/Test_normal.pdf`
3. NIST/SEMATECH *e-Handbook of Statistical Methods*, `http://www.itl.nist.gov/div898/handbook/`
4. PROPHET StatGuide *Examining normality test results*, `http://www.basic.northwestern.edu/statguidefiles/n-dist_exam_res.html`
5. Saporta, G., *Probabilités, Analyse des données et Statistique*, Technip, 2ème édition, 2006.
6. Sneyers, R., *Sur les tests de normalité*, in Revue de Statistique Appliquée, Tome 22, n.22, 1974, `http://archive.numdam.org/ARCHIVE/RSA/RSA_1974__22_2/RSA_1974__22_2_29_0/RSA_1974__22_2_29_0.pdf`.
7. Thode Jr., H.C., *Testing for Normality*, Marcel Dekker, New York, 2002.