Le test de normalité de D'Agostino-Pearson

Il s'agit d'un test de normalité adapté aux variables comprenant de nombreux ex aequo.

Position du problème.

Les tests classiques de normalité, tels que Kolmogorov-Smirnov, Lilliefors, Shapiro-Wilk, etc sont très sensibles à la présence d'ex aequo.

Considérons par exemple une variable statistique obtenue à partir d'une échelle de Likert en 5 ou 7 points.

Si on estime que l'échelle de Likert fournit une variable purement ordinale, les questions de normalité ne se posent pas, car il n'y a pas lieu de chercher une approximation de cette variable par une loi continue. Mais il n'est alors pas pertinent non plus de chercher à définir une variable donnant un score de synthèse en calculant une somme ou une moyenne de telles variables.

En revanche, on peut aussi juger que l'échelle de Likert fournit une mesure *relativement imprécise* d'une variable continue. Ainsi :
$X = 1$ correspond en fait à $X \leq 1{,}5$
$X = 2$ correspond en fait à $1{,}5 < X \leq 2{,}5$
etc.

Dans ce cas, le problème de la normalité de X dans la population parente a un sens. Mais les tests tels que Kolmogorov-Smirnov, Lilliefors, Shapiro-Wilk, Anderson-Darling ont l'inconvénient d'être très sensibles à la présence d'ex aequo, comme le montrent les manipulations ci-dessous.

Exemple 1 :
X est la série statistique des réponses de 48 sujets sur une échelle de Likert en 5 points :
X : 3 5 5 3 2 4 5 2 3 1 2 1 2 4 2 2 3 3 3 4 3 3 4 5 4 3 4 3 2 1 3 5 3 4 3 1 3 2 3 4 3 2 1 3 2 1 3 2

Seul le test de Kolmogorov-Smirnov (dont l'usage n'est pas véritablement correct ici) conduit à une p-value supérieure à 5%. Les autres tests concluent tous sur un défaut de normalité.

```
>library(nortest)

> ks.test(X,"pnorm",mean=mean(X),sd=sd(as.vector(X)))
      One-sample Kolmogorov-Smirnov test
data:  X
D = 0.1932, p-value = 0.05562
alternative hypothesis: two-sided
Message d'avis :
In ks.test(X, "pnorm", mean = mean(X), sd = sd(as.vector(X))) :
  impossible de calculer les p-values correctes avec des ex-aequos

> lillie.test(X)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  X
D = 0.1932, p-value = 0.0001105

> shapiro.test(X)
      Shapiro-Wilk normality test
data:  X
W = 0.914, p-value = 0.001842

> ad.test(X)
      Anderson-Darling normality test
data:  X
```

```
A = 1.6241, p-value = 0.0003092
```

En revanche, si on ajoute à chacune des 48 valeurs entières de la série statistique X une partie décimale aléatoire comprise entre -0,5 et 0,5 (X représente alors l'arrondi entier des valeurs X1 générées), ces différents tests concluent sur la normalité de la série ainsi obtenue :

```
> X1 <- X+runif(48)-.5

> ks.test(X1,"pnorm",mean=mean(X1),sd=sd(as.vector(X1)))

      One-sample Kolmogorov-Smirnov test
data:  X1
D = 0.0709, p-value = 0.955
alternative hypothesis: two-sided

> lillie.test(X1)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  X1
D = 0.0709, p-value = 0.7893

> shapiro.test(X1)
      Shapiro-Wilk normality test
data:  X1
W = 0.9772, p-value = 0.4677

> ad.test(X1)
      Anderson-Darling normality test
data:  X1
A = 0.2523, p-value = 0.7232
```

Exemple 2 :

A l'inverse, on génère une série statistique de 48 valeurs tirées d'une loi normale dont les paramètres sont ceux de la variable X ci-dessus.

```
> Y <- rnorm(48, mean=mean(X),sd=sd(as.vector(X)))

> Y
 [1] 2.3178870 3.7099003 1.4615044 1.7937996 2.8787223 3.6629975 3.0464251 2.2819286
2.6065499 3.4581155 2.8317473 4.6929409 1.0392315 4.0965538
[15] 2.7516116 2.1332957 4.1547208 2.1597656 3.3326866 0.4377083 3.6537238 3.6942626
2.6554317 3.4626125 4.3065276 3.1518447 0.4154250 3.5626645
[29] 3.2156494 0.9715125 3.9933763 2.7724458 2.8450445 2.7503986 1.0119739 4.2808468
2.5277786 3.4208558 3.6717145 2.5735135 2.6232302 3.1563150
[43] 2.6783873 1.4663685 4.7654520 1.7376890 2.8591073 4.7457353
```

Sans surprise, aucun des 4 tests précédents ne met en évidence de défaut de normalité de cette série.

```
> ks.test(Y, "pnorm",mean=mean(Y), sd=sd(Y))
      One-sample Kolmogorov-Smirnov test
data:  Y
D = 0.1047, p-value = 0.6307
alternative hypothesis: two-sided

> lillie.test(Y)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  Y
D = 0.1047, p-value = 0.2089

> shapiro.test(Y)
      Shapiro-Wilk normality test
data:  Y
W = 0.9696, p-value = 0.2447
```

```
> ad.test(Y)
      Anderson-Darling normality test
data:  Y
A = 0.4167, p-value = 0.319
```

Formons alors la série des arrondis entiers des 48 valeurs précédentes.

```
> Y1 <- as.integer(Y+.5)
> Y1
 [1] 2 4 1 2 3 4 3 2 3 3 3 5 1 4 3 2 4 2 3 0 4 4 3 3 4 3 0 4 3 1 4 3 3 3 1 4 3 3 4 3
3 3 3 1 5 2 3 5
```

Les quatre tests précédents indiquent alors un défaut de normalité de Y1

```
> ks.test(Y1, "pnorm",mean=mean(Y1), sd=sd(Y1))
      One-sample Kolmogorov-Smirnov test
data:  Y1
D = 0.2641, p-value = 0.002477
alternative hypothesis: two-sided
Message d'avis :
In ks.test(Y1, "pnorm", mean = mean(Y1), sd = sd(Y1)) :
  impossible de calculer les p-values correctes avec des ex-aequos

> lillie.test(Y1)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  Y1
D = 0.2641, p-value = 4.872e-09

> shapiro.test(Y1)
      Shapiro-Wilk normality test
data:  Y1
W = 0.9018, p-value = 0.0007196

> ad.test(Y1)
      Anderson-Darling normality test
data:  Y1
A = 2.2048, p-value = 1.109e-05
```

N.B. Les séries statistiques utilisées dans les exemples ci-dessus sont enregistrées dans la feuille de données Dago-Ex1.csv.

2) Le test de D'Agostino - Pearson

Le test utilise la forme de la distribution de X, telle qu'elle est mesurée par l'asymétrie (skewness) et l'aplatissement (kurtosis).
Ces paramètres de forme se calculent à partir des moments centrés d'ordre 3 et 4 :

$$\mu_3 = \frac{\sum (X_i - \mu)^3}{n} \quad ; \quad \mu_4 = \frac{\sum (X_i - \mu)^4}{n}$$

Pour une variable définie sur une population, l'asymétrie et l'aplatissement sont définis par :

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \quad ; \quad \beta_2 = \frac{\mu_4}{\sigma^4} \quad ; \quad \gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Pour une variable distribuée selon une loi normale, on a : $\gamma_1 = 0$ et $\beta_2 = 3$.

Lorsque X représente les valeurs de la variable dépendante sur un échantillon de taille n, les estimations de l'asymétrie et de l'aplatissement de la variable dépendante dans la population parente (estimateurs sans biais) sont données par :

$$g_1 = \frac{m_3}{\tilde{s}^3}$$

où m_3 est l'estimation du moment d'ordre 3 : $m_3 = \dfrac{n\sum\left(x_i-\bar{x}\right)^3}{(n-1)(n-2)}$ et $\tilde{s}$ est l'écart type corrigé : $\tilde{s} = \left(\dfrac{\sum\left(x_i-\bar{x}\right)^2}{n-1}\right)^{1/2}$.

$$g_2 = \frac{k_4}{\tilde{s}^4}$$

où k_4 (qui n'est en fait pas une estimation du moment d'ordre 4) est calculé par :
$k_4 = \dfrac{n(n+1)\sum\left(x_i-\bar{x}\right)^4}{(n-1)(n-2)(n-3)} - 3\dfrac{\left(\sum\left(x_i-\bar{x}\right)^2\right)^2}{(n-2)(n-3)}$ et $\tilde{s}$ est l'écart type corrigé : $\tilde{s} = \left(\dfrac{\sum\left(x_i-\bar{x}\right)^2}{n-1}\right)^{1/2}$.

Test de l'asymétrie d'une distribution

Les hypothèses du test sont alors :
H$_0$: $\gamma_1 = 0$
H$_1$: $\gamma_1 \neq 0$

Tabachnick et Fidell ont montré que l'erreur standard sur l'asymétrie peut être estimée par : $SE = \sqrt{\dfrac{6}{n}}$. On peut utiliser comme statistique de test la valeur : $z = \dfrac{g_1}{SE}$. Mais cette statistique ne suit pas une loi mathématique classique, et il est nécessaire de recourir à des tables spécifiques. De telles tables sont disponibles par exemple aux adresses :
http://nte-serveur.univ-lyon1.fr/nte/immediato/math2001/tables/asymetrie.htm
http://mvpprograms.com/help/mvpstats/distributions/SkewnessCriticalValues

D'Agostino et Zar proposent de calculer à partir de g_1 une statistique qui suit approximativement une loi normale centrée réduite.

On calcule successivement :

$$A = \frac{(n-2)g_1}{\sqrt{n(n-1}}\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$

$$B = \frac{3\left(n^2+27n-70\right)\left(n+1\right)\left(n+3\right)}{\left(n-2\right)\left(n+5\right)\left(n+7\right)\left(n+9\right)}$$

$$C = \sqrt{2\left(B-1\right)}-1$$

$$D = \sqrt{C}$$

$$E = \frac{1}{\sqrt{\ln\ D}}$$

$$F = \frac{A}{\sqrt{\dfrac{2}{C-1}}}$$

$$z_{g_1} = E\ \ \ln\left(F+\sqrt{F^2+1}\right)$$

Sous H$_0$, cette dernière statistique suit une loi normale centrée réduite.

Test de l'aplatissement d'une distribution

Les hypothèses du test sont alors :

H$_0$: $\gamma_2 = 0$
H$_1$: $\gamma_2 \neq 0$

Il a été montré que l'erreur type sur l'aplatissement peut être estimé par par : $SE = \sqrt{\frac{24}{n}}$. On peut utiliser comme statistique de test la valeur : $z = \frac{g_2}{SE}$ à condition de recourir à des tables spécifiques. De telles tables sont disponibles par exemple aux adresses :
http://nte-serveur.univ-lyon1.fr/nte/immediato/math2001/tables/aplatissement.htm
http://mvpprograms.com/help/mvpstats/distributions/KurtosisCriticalValues

Comme dans le cas de l'asymétrie, D'Agostino et Zar proposent de calculer une statistique qui suit approximativement une loi normale centrée réduite.

On calcule successivement :

$$G = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

$$H = \frac{(n-2)(n-3)g_2}{(n+1)(n-1)\sqrt{G}}$$

$$J = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$K = 6 + \frac{8}{J}\left(\frac{2}{J} + \sqrt{1+\frac{4}{J^2}}\right)$$

$$L = \frac{1-\frac{2}{K}}{1+H\sqrt{\frac{2}{K-4}}}$$

$$z_{g_2} = \frac{1-\frac{2}{9K}-\sqrt[3]{L}}{\sqrt{\frac{2}{9K}}}$$

Sous H$_0$, cette dernière statistique suit une loi normale centrée réduite.

Remarque. Dans l'ouvrage de Sheskin, c'est la valeur absolue de g_2 et non sa valeur algébrique qui est utilisée. En revanche, les équations ci-dessus correspondent aux calculs faits par la procédure dagoTest du package fBasics de R.

Le test de normalité de D'Agostino

L'hypothèse nulle est alors la normalité de la distribution parente, l'hypothèse alternative est le défaut de normalité.

La statistique de test (souvent notée K^2) proposée par D'Agostino est :

$$\chi^2 = z_{g_1}^2 + z_{g_2}^2$$

Sous H$_0$, cette statistique suit approximativement une loi du khi-2 à 2 degrés de liberté.

Une critique faite au test de D'Agostino : les statistiques z_{g_1} et z_{g_2} ne sont pas indépendantes. La loi suivie par la statistique de D'Agostino n'est donc qu'approximativement un khi-2.
Pour appliquer ce test, on exige généralement que l'effectif de l'échantillon soit supérieur ou égal à 20.

Les différents calculs décrits ci-dessus sont réalisés, pour la série statistique X et pour deux autres séries données par Sheskin, dans le classeur Excel Dagostino.xls

Le test de normalité de D'Agostino avec le package fBasics de R

Le test de D'Agostino est disponible dans le package fBasics de R. Ce test, appliqué aux séries X, X1, Y, Y1 précédentes, produit les résultats suivants.

X est la série de valeurs entières comprises entre 1 et 5 fournies par une échelle de Likert soumise à 48 sujets.

```
> library(fBasics)
> dagoTest(X)
Title:
 D'Agostino Normality Test
Test Results:
  STATISTIC:
    Chi2 | Omnibus: 0.8705
    Z3  | Skewness: 0.3799
    Z4  | Kurtosis: -0.8521
  P VALUE:
    Omnibus  Test: 0.6471
    Skewness Test: 0.704
    Kurtosis Test: 0.3941
Description:
 Wed Dec  7 17:59:15 2011 by user:
```

Ainsi, la procédure calcule les statistiques sur l'asymétrie et l'aplatissement (nommées ici Z3 et Z4), puis la statistique de D'Agostino, nommée Chi2 omnibus.
On voit que, contrairement aux tests utilisés en introduction, le test de D'Agostino permet de conclure à la normalité de la distribution parente de l'échantillon.

Pour la série X1, obtenue en introduisant des décimales aléatoires à la série X, on obtient :

```
> dagoTest(X1)
Title:
 D'Agostino Normality Test
Test Results:
  STATISTIC:
    Chi2 | Omnibus: 1.7505
    Z3  | Skewness: -0.2867
    Z4  | Kurtosis: -1.2916
  P VALUE:
    Omnibus  Test: 0.4168
    Skewness Test: 0.7743
    Kurtosis Test: 0.1965
Description:
 Wed Dec  7 18:00:27 2011 by user:
```

La normalité est aussi acceptée, en accord cette fois avec les autres tests de normalité.

La série Y est une série de 48 valeurs aléatoires générées selon une loi normale. Sans surprise, le test de D'Agostino conclut sur la normalité de cette série :

```
> library(fBasics)
> dagoTest(Y)
Title:
 D'Agostino Normality Test
Test Results:
  STATISTIC:
    Chi2 | Omnibus: 1.5632
```

```
    Z3  | Skewness: -1.2496
    Z4  | Kurtosis: -0.0421
  P VALUE:
    Omnibus  Test: 0.4577
    Skewness Test: 0.2114
    Kurtosis Test: 0.9664
Description:
 Tue Dec  6 11:01:43 2011 by user:
```

Enfin, la série Y1 est formée des arrondis entiers de la série précédente. On obtient alors :

```
> dagoTest(Y1)
Title:
 D'Agostino Normality Test
Test Results:
  STATISTIC:
    Chi2 | Omnibus: 3.2229
    Z3  | Skewness: -1.7038
    Z4  | Kurtosis: 0.5658
  P VALUE:
    Omnibus  Test: 0.1996
    Skewness Test: 0.08843
    Kurtosis Test: 0.5715
Description:
 Tue Dec  6 11:01:50 2011 by user:
```

Alors que les autres tests concluaient sur un défaut de normalité de cette série, le test de D'Agostino conclut en sens inverse.

Conclusion

Les exemples ci-dessus illustrent la sensibilité des tests usuels de normalité (Lilliefors, Shapiro-Wilk, etc) aux ex aequo et l'intérêt du test de D'Agostino sur de telles données. On pourrait penser, au vu de ces exemples, que le test de D'Agostino est très conservatif, et peu puissant. Mais plusieurs auteurs indiquent au contraire que ce test est plus puissant que les tests classiques.

On pourra noter également que d'autres tests de normalité ont été proposés, notamment le test de Jarque-Bera. Un exposé plus détaillé pourra être trouvé à l'adresse :
eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf

Bibliographie.

D'Agostino, R.B. (1986), Tests for the Normal Distribution, in D'Agostino, R.B. and Stephens, M.A. (eds.), Goodness-of-fit Techniques (pp. 367-419). New York: Marcel Dekker.

D'Agostino, R.B., Pearson, E.S. (1973), Tests of Departure of Normality. Empirical Results for the distribution of b2 and $\sqrt{b1}$, Biometrika, 60, pp 613-622

Sheskin, David J., Handbook of Parametric and nonparametric statistical procedures, 3rd edition, Chapman & Hall, 2004.

Tabachnick, B.G., Fidell, L.S., Using Multivariate Statistics, 4th edition, Boston: Allyn & Bacon