Revue de statistique appliquée

F. Daumas

Méthodes de normalisation des données

Revue de statistique appliquée, tome 30, n° 4 (1982), p. 23-38
<http://www.numdam.org/item?id=RSA_1982__30_4_23_0>

METHODES DE NORMALISATION DES DONNEES

F. DAUMAS

Centre National Universitaire Sud de Calcul

Le but de cet article est de réaliser un survol synthétique des principales techniques de normalisation des données présentes dans la littérature statistique. Celles-ci sont présentées hors du cadre des modèles linéaires où elles sont souvent utilisées. Un intérêt particulier est porté aux transformations de BOX et COX.

INTERET DE LA NORMALISATION

Les statistiques paramétriques classiques étant essentiellement basées sur la loi normale, assumer la normalité pour une variable c'est accéder à un large éventail d'outils statistiques. Malheureusement les phénomènes naturels ne sont pas en général normalement distribués. Aussi le statisticien pourra avoir trois tentations :

– La première tentation consiste à supposer a-priori que la population est normale. Souvent des distributions observées tendent vers une loi normale lorsque le nombre n des observations tend vers l'infini, il est alors suffisant pour n assez grand d'assumer la normalité. Pour des valeurs plus faibles de n cette approximation peut se révéler grossière et soulève alors le problème de la robustesse des résultats obtenus. Dans quelle mesure ces résultats sont-ils encore corrects lorsque l'hypothèse de normalité n'est pas vérifiée ?

– La seconde tentation consiste à abandonner l'objectif de la normalité des données et de chercher des statistiques valables pour l'ensemble des distributions continues (distribution free) ou de s'accomoder des propriétés intrinsèques de la loi mise en évidence.

– La troisième tentation est celle qui sera développée dans la suite, elle consiste à essayer de se ramener au moyen de transformations des données à des distributions normales. En fait la qualité essentielle de la loi normale étant sa symétrie sphérique, le seul fait de symétriser les données présente un réel progrés.

Mots-Clés : **Normalité. Transformation de données. Transformations normalisatrices. Transformations de Box et Cox.**

Solution théorique

Lorsqu'on choisit de normaliser les données encore faut-il savoir quelle transformation employer. Si la distribution théorique F de la variable observée X est connue (ce qui est souvent difficile dans le contexte expérimental) la transformation Z.F, où Z est l'inverse de la fonction de répartition de la loi normale centrée réduite, permet de normaliser les données.

Ainsi, si X, définie sur la demi droite réelle positive, suit une loi log-normale de moyenne m est d'écart type s, $Y = \log(X)$ suit une loi normale.

La transformation Z.F conserve évidemment le rang des observations mais son application pratique se heurte à la difficulté de connaître F de façon précise à partir de l'expérimentation. Substituer à la solution exacte des transformations qui permettent d'approcher au mieux la normalité semble alors une démarche naturelle. Même lorsque F est connu il est souvent utile de trouver de telles transformations.

Techniques particulières de normalisation de certaines lois

On donne ici deux exemples de ces techniques.

Loi Gamma 23) : Soit X une variable qui suit une loi Gamma de paramètres a et b, en particulier : si $a = 1$, X suit une loi exponentielle de paramètre $1/b$; si $a = n/2$ et $b = 2$, X suit une loi du khi-2 à n degrés de libertés.

La transformation racine carrée qui stabilise la variance, permet d'obtenir pour a assez grand, une distribution approximativement normale de moyenne $\sqrt{b(a - 1/4)}$ et de variance $b/4$. Cependant WILSON et HILFERTY [33] ont montré en étudiant les coefficients d'asymétrie b1 et d'aplatissement $\gamma 2 = b2 - 3$ que la transformation racine cubique donne une approche plus précise de la normalité. Mais cette transformation ne stabilise pas la variance.

La moyenne est alors : $(1 - 1/9a) . (ab)^{1/3}$ et la variance : $(b^2/a)^{1/3} . (1/9)$.

Loi inverse de Gauss (32) : Soit X une variable positive qui suit une loi inverse de Gauss de moyenne $\mu > 0$ et de paramètre de concentration $a > 0$. En particulier $Y = X/\mu$ suit une distribution de Wald de paramètre $\theta = a/\mu$. Whitmore et Yalovsky (32) montrent que la transformation $W = \sqrt{\theta} . \log(Y) + 1/(2 . \sqrt{\theta})$ approche très vite la normalité. W est un peu plus concentré autour du mode que la loi normale.

Normalisation par des méthodes indirectes

Certaines transformations qui n'ont pas pour but principal de normaliser les données ont souvent pour effet secondaire de rendre les données plus conformes au modèle normal. C'est souvent le cas lorsqu'on désire construire une relation linéaire entre un prédictant Y et un groupe de prédicteurs X1 , X2 , . . . , Xp. Les hypothèses requises pour la validité du modèle ne sont pas toujours vérifiées. On est alors courramment amené à essayer de trouver un changement d'échelle pour mesurer Y (c.a.d. une transformation non linéaire de Y) qui permette :

- Soit de normaliser la distribution des erreurs ;
- Soit de stabiliser leur variance ;
- Soit de rendre les effets additifs en éliminant les interactions.

Dans la pratique on remarque que la recherche d'un de ces objectifs augmente les chances de réaliser les autres, sans que ce soit la règle.

Transformation visant à stabiliser la variance : Pour construire de telles transformations il suffit d'obtenir une relation, théorique ou empirique, entre la variance et la moyenne. Cette 'épendance entre variance et moyenne s'accompagne souvent d'une dissymétrie excessive des données qui tend donc à disparaître avec la transformation qui stabilise la variance. Le tableau (I) donné en annexe 1 est tiré d'une publication de BARTLETT [6], il résume pour un certain nombre de relations entre la moyenne et la variance d'une statistique t les transformations stabilisatrices correspondantes.

Transformations visant à éliminer les interactions : Ces transformations le plus souvent suggérées par l'analyse des résidus consistent pratiquement à minimiser la valeur du rapport de la somme des carrés des différences inter groupes sur la somme des carrés des résidus, pour une classe de transformations. Relativement peu de travaux ont été réalisés dans ce domaine.

Il est en outre possible, lorsqu'une fréquence est prise comme variable dans une analyse, d'utiliser les probits pour obtenir l'additivité des effets [6]. Cette transformation est alors évidemment normalisatrice, mais elle ne stabilise pas la variance.

Aide au choix de transformations normalisatrices

Certaines techniques souvent à base graphique permettent d'orienter la recherche de transformations susceptibles de normaliser les données, en mettant en évidence sur celles-ci les écarts à la normalité (le plus souvent les écarts à la symétrie). Si le tracé d'histogramme est la technique la plus connue d'autre tout aussi simples peuvent être utilisées avec profit (voir GNANADESIKAN [16]).

Analyse des résidus : Dans le cadre des modèles linéaires le tracé graphique des points ayant pour abscisses les valeurs ajustées et pour ordonnées les résidus correspondants donne une idée grossière :

- de l'homoscédasticité, en observant la dispersion des résidus pour différents sous ensembles de valeurs ajustées ;
- de la non normalité en observant la dissymétrie de la dispersion des résidus ;
- de la non additivité des effets en mettant en évidence l'existence d'une régression non linéaire des résidus par rapport aux valeurs ajustées.

Dans tous les cas ces observations peuvent suggérer une transformation appropriée de la variable observée. Voir sur ce sujet les travaux d'ANSCOMBE et TUKEY [4].

Q – Q Plots (16) : Si X et Y sont deux variables de distributions respectives F et G, le principe des Q – Q plots est de porter dans un schéma les points dont les coordonnées sont les images de $p \in [0,1]$ par les fonctions inverses de F et G, dans le but de comparer F et G.

– si les deux lois coïncident les points s'alignent suivant la première bissectrice.
– si les deux lois diffèrent par leur localisation l'alignement des points se fait suivant une droite de pente 1 qui coupe l'axe horizontal en un point différent de l'origine.
– si les deux lois diffèrent par leur échelle l'alignement des points se fait suivant une droite de pente différente de 1.

Ainsi si on prend pour F la distribution théorique de la loi normale centrée réduite et si G est la distribution empirique de la variable Y centrée réduite, les Q – Q plots permettront d'observer l'écart à l'hypothèse F = G, représentée sur le schéma par la première bissectrice. Cette observation pourra alors inspirer un changement de coordonnées permettant de se rapprocher de l'hypothèse. Une courbure parabolique suscitera, par exemple, une transformation racine carrée.

Probits : Soit f la fréquence avec laquelle la variable X a pris la valeur x, on construit la valeur $y = 5 + Z(f)$*. Si X suit une loi normale, le nuage des points (x , y) doit théoriquement s'aligner le long d'une droite. Celle-ci a pour pente l'inverse de l'écart type et passe par le point qui a pour abscisse la moyenne et pour ordonnée la valeur 5. Toute courbure particulière du nuage de points traduit alors de même que dans le paragraphe précédent un écart à la normalité de X et peut aider au choix d'un changement de coordonnées.

Ces techniques décrites ici pour aider dans le choix d'une transformation, peuvent évidemment servir, une fois la transformation faite, à constater la qualité de l'effet normalisateur.

Transformations polynomiales

Le principe est de chercher les coefficients d'un polynôme P tel que la fonction polynôme associée permette de normaliser les données. La résolution de ce problème, qui fait intervenir le développement d'EDGEWORTH d'une densité, est due à CORNISH et FISCHER (voir STUART et KENDALL [22]). Cependant la fonction P n'étant pas toujours croissante, ces transformations ne conservent pas en général le rang des observations. Ceci enlève énormément d'intérêt à ces transformations.

Transformations de Johnson

Le problème de l'identification de la loi d'une variable, a conduit de nombreux auteurs à essayer de construire un système permettant de décrire en termes mathématiques toute la gamme des distributions possibles (voir sur ce sujet les travaux de PEARSON, EDGEWORTH, GRAM-CHARLIER [21] et [22]). En particulier JOHNSON a cherché à identifier la distribution d'une variable au moyen de la transformation à lui appliquer pour obtenir une distribution normale. Elargissant l'optique du paragraphe précédent, puisqu'on ne se limite plus aux seules transformations polynomiales, on ne peut obtenir qu'une meilleure approche de la normalité.

JOHNSON considère les transformations du type suivant :

$$t : x \longrightarrow t(x) = a + b \cdot g((x - c)/d)$$

---

* Z est défini comme vu plus haut.

– a, b, c et d sont des paramètres réels ;
– g est une fonction monotone ne dépendant d'aucun paramètre variable.

Sans perte de généralité g peut être supposée croissante et b et d positifs.

JOHNSON établit alors une classification des principales lois en trois systèmes associés chacun à une transformation g particulière et donc à une famille de transformations t.

– Si $g(z) = \log(z)$ le système associé est noté Sl (lois log-normales).
– Si $g(z) = \log(z/(1 - z))$ le système associé est noté Sb.
– Si $g(z) = \text{Argsh}(z)$ le système associé est noté Su.

L'estimation des paramètres a, b, c, d des transformations de JOHNSON s'effectue généralement à l'aide de l'expression des moments de la variable $Z = (X - c)/d$. JOHNSON et KOTZ [21] donnent pour chacun des systèmes l'expression des moments non centrés d'ordre r de Z, ainsi que des indications pour le calcul des paramètres (voir aussi pour les lois Su l'article de WINTERBON [34]). Pour le système Sb l'expression des moments ne pouvant se mettre sous forme utilisable il est conseillé d'utiliser les quantiles pour estimer les paramètres.

Ces transformations conservent le rang des observations. Utiliser les transformations de JOHNSON pour normaliser des données suppose que l'on connaisse le système auquel appartient la distribution observée. JOHNSON et KOTZ [21] donnent un moyen pour déterminer graphiquement, à partir des valeurs calculées des coefficients d'asymétrie et d'aplatissement, le type de loi suivie par les données.

Transformations de Haldane

Le paragraphe précédent montre que le calcul des coefficients d'asymétrie et d'aplatissement permet d'envisager l'utilisation d'une transformation de JOHNSON. Dans le même ordre d'idée ce qui suit montre qu'il est possible lorsqu'on connaît les cumulants de déterminer une transformation normalisatrice. Si la variable observée X a tous ses cumulants Kr de l'ordre de n (n étant la taille de l'échantillon) alors HALDANE [17] démontre qu'on obtient :

– un coefficient d'asymétrie de l'ordre de : $n^{-3/2}$ si on considère la transformation :

$$x \longrightarrow y = (1 + (x - K1)/K1)^{k}$$

avec :

$$k = 1 - K1 \,.\, K3/3(K2)^2$$

– en plus un coefficient d'aplatissement de l'ordre de : n^{-2} si on considère :

$$x \longrightarrow z = (1 + (x - K1)/t)^{h}$$

avec :

$$h = (16\,K3^2 - 9\,K4 \,.\, K2)/(20\,K3^2 - 9\,K4 \,.\, K2)$$

et

$$t = 12\,K3 \,.\, K2^2/(20\,K3^2 - 9\,K4 \,.\, K2)$$

On trouve en particulier (22) pour la loi du khi2 : $k = 1/3$ (en accord avec WILSON et HILFERTY) et $h = -5/3$.

Transformations puissances

Définition 1 : On appelle transformations puissances la famille des transformations de paramètres a, b, c, d réels définies pour tout x supérieur à $-$ c par :

$$x \longrightarrow a + b\,(x + c)^d$$

Une sous famille particulière est obtenue en prenant $c = k2$ et $b = 1/d = -a = 1/k1$ pour k1 différent de 0. Si k désigne alors le vecteur (k1 , k2) on note pour k1 différent de zéro et x supérieur à $-$ k2 :

$$g_k(x) = \frac{(x + k2)^{k1} - 1}{k1}$$

On montre facilement que pour x fixé supérieur à $-$ k2 la limite, lorsque k1 tend vers zéro, de la fonction précédente est égale à $\log(x + k2)$. Ceci conduit donc à considérer les transformations suivantes :

Définition 2 : On appelle transformations de BOX et COX (9) la famille des transformations de paramètre $k = (k1, k2)$ appartenant à $R \times R$ définies pour x supérieur à $-$ k2 par :

$$g_k(x) = \begin{cases} \log(x + k2) & \text{si } k1 = 0 \\ \dfrac{(x + k2)^{k1} - 1}{k1} & \text{sinon} \end{cases}$$

Dans la littérature $g_k(x)$ est souvent noté : $x^{(k)}$

Lorsque X est une variable réelle strictement positive, il est suffisant de considérer la famille des transformations à un seul paramètre réel : k, définies en prenant $k2 = 0$ et $k1 = k$ dans l'expression précédente.

Propriétés des transformations de BOX et COX

i) L'application $g : (k1\ , k2\ , x) \longrightarrow g_k(x)$ est :

continue et dérivable en k1.

ii) L'application $g_k : x \longrightarrow g_k(x)$ est :

dérivable sur $]-k2\ , +\infty[$ de dérivée $(x + k2)^{k1 - 1}$

Conséquences

$-$ Si k1 est très proche d'une valeur donnée h (par exemple 0 ou 1/2) il sera possible d'utiliser la transformation de paramètre (h , k2) en lieu et place de la transformation de paramètre (k1 , k2) avec la certitude que pour tout x fixé les deux quantités obtenues par ces deux transformations seront très proches.

– La transformation de paramètre k est strictement croissante sur $]-k_2, +\infty[$ à valeurs dans l'intervalle $]A, B[$. Avec :

$$\begin{array}{lll} \text{si} & k1 = 0, & A = -\infty \text{ et } B = +\infty \\ \text{si} & k1 > 0, & A = -1/k1 \text{ et } B = +\infty \\ \text{si} & k1 < 0, & A = -\infty \text{ et } B = -1/k1 \end{array}$$

Elle conserve donc le rang des observations.

– La transformation de paramètre k admet une fonction réciproque strictement croissante sur $]A, B[$ à valeur dans $]-k2, +\infty[$.

$$g_k^{-1}(y) = \begin{cases} \exp(y) - k2 & \text{si } k1 = 0 \\ (1 + k1 . y)^{1/k1} - k2 & \text{sinon} \end{cases}$$

– La transformation de paramètre k est convexe (respectivement concave) lorsque $k1 > 1$ (respectivement $k1 < 1$). Ainsi : pour k1 inférieur à 1 (respectivement supérieur) la moyenne μ de Y est inférieure (resp. supérieure) à la transformée de la moyenne de X, et donc la médiane de X est inférieure (resp. supérieure) à la moyenne. La connaissance des positions relatives de la médiane et de la moyenne des données initiales permet de situer la valeur du paramètre k1 par rapport à 1.

– Il est possible ([19] et [22]) d'obtenir une relation entre les moments non centrés d'ordre r de X + k2 et les moments centrés de Y, en prenant l'espérance mathématique du développement en série entière, au voisinage de $y = \mu$, de :

$$(k2 + g_k^{-1}(y))^r$$

Si Y est supposée normalement distribuée l'expression de ses moments centrés est connue en fonction de σ. Alors lorsque $k1 = 0$ on retrouve la relation connue pour la loi log-normale.

Estimation du paramètre de normalisation k : Si le paramètre de normalisation existe, son estimation s'effectue souvent, en pratique, par la méthode du maximum de vraisemblance. Celle-ci n'exige aucune connaissance préalable des données. Il existe dans la littérature d'autres possibilités d'estimer k (voir en particulier BOX et COX [9]).

Soit $X1, X2, \ldots, Xn$ un n-échantillon de variable parente X et $Y1, Y2, \ldots, Yn$ les observations supposées normalement distribuées, de moyenne μ et d'écart type σ, obtenues aprés une transformation puissance de paramètre k. Le logarithme de la fonction de vraisemblance a alors pour expression :

$$L(k) = (k1 - 1) \sum_{i=1}^{n} \log(Xi + k2) - (1/2) \sum_{i=1}^{n} ((g_k(Xi) - \mu)/\sigma)^2$$
$$- (n/2) \log(2\pi\sigma^2)$$

BOX et COX proposent alors de maximiser cette expression par rapport à μ et σ. Cela revient à remplacer μ et le carré de σ par les estimateurs du maximum de vraisemblance de la moyenne et de la variance des données transformées. On obtient alors à une constante additive prés :

$$Lm(k) = (k1 - 1) \sum_{i=1}^{n} \log(Xi + k2) - n \log(\hat{\sigma}_k)$$

Proposition : L'estimateur du maximum de vraisemblance $\hat{k}$ de k est la valeur qui maximise l'expression de Lm(k).

L'approximation de $\hat{k}$ s'effectue souvent, dans la pratique, en visualisant, au moyen du graphe de la fonction Lm, la valeur correspondant au maximum, ainsi que la région de confiance approximative associée à un seuil $1 - \alpha$. Cette dernière correspond aux valeurs de k qui vérifient :

$$2(\mathrm{Lm}(\hat{k}) - \mathrm{Lm}(k)) < \chi_1^2(\alpha)$$

où le second membre désigne le quantile correspondant à α d'un khi2 à un seul degré de liberté. Ceci résulte des propriétés asymptotiques du rapport de vraisemblance.

Des méthodes numériques permettent une détermination plus rigoureuse de $\hat{k}$, elles exigent la connaissance des dérivées de Lm.

Proposition :

$$\frac{d\,\mathrm{Lm}(k)}{dk1} = \sum_{i=1}^{n} \log(Xi + k2) - n\,\frac{(\underset{*}{Y} - \hat{\mu}_k 1)'(\underset{*}{u}_k - \hat{\eta}_k 1)}{\|\underset{*}{Y} - \hat{\mu}_k 1\|^2} + (n/k1)$$

$$\frac{d\,\mathrm{Lm}(k)}{dk2} = (k1 - 1)\sum_{i=1}^{n} (Xi + k2)^{-1} - n\,\frac{(\underset{*}{Y} - \hat{\mu}_k 1)'(\underset{*}{v}_k - \hat{w}_k 1)}{\|\underset{*}{Y} - \hat{\mu}_k 1\|^2}$$

où :

$$\hat{\mu}_k = (1/n)\sum_{i=1}^{n} g_k(Xi)\ ;\quad \hat{\eta}_k = (1/n)\sum_{i=1}^{n} u_i(k)\ ;\quad \hat{w}_k = (1/n)\sum_{i=1}^{n} v_i(k)$$

avec

$$u_i(k) = \frac{(Xi + k2)^{k1}\log(Xi + k2)}{k1}\ ;\quad v_i(k) = (Xi + k2)^{k1-1}$$

$\underset{*}{Y}$, $\underset{*}{u}_k$, $\underset{*}{v}_k$ sont les vecteurs colonnes qui ont respectivement pour composantes Yi, $u_i(k)$, $v_i(k)$. 1 est la matrice colonne $n \times 1$ dont tous les éléments sont égaux à 1.

Validité de la normalisation : La détermination d'un $\hat{k}$ ne signifie pas obligatoirement que les données initiales puissent être normalisées au moyen de la transformation correspondante. Il est donc logique et nécessaire de répondre à la question : les données transformées par la transformation puissance de paramètre égal à $\hat{k}$ peuvent-elles être raisonnablement considérées comme normalement distribuées ? . Il semble souhaitable que le test choisi prenne plus en compte les différences au niveau des points fortement probables que celles constatées pour les points de faible probabilité.

De ce point de vue, il peut être intéressant d'utiliser le test basé sur la distance de HELLINGER entre deux lois (voir (27)) :

$$d(f,g) = \left(\int (\sqrt{f(x)} - \sqrt{g(x)})^2\,dx\right)^{1/2}$$

Le principe de ce test est de discrétiser en considérant q classes d'observations. Si la variable observée X a ; Fn pour distribution empirique sur l'échantillon de taille n considéré et F pour distribution théorique, et si G est une distribution théorique que l'on désire ajuster, le problème consiste à tester l'hypothèse H : F = G. K. MATUSITA [27] a montré que pour tout réel positif fixé η et pour tout ϵ donné on a :

– $d^2(G, Fn) < (q - 1)/(n\epsilon)$ avec une probabilité supérieure ou égale à $1 - \epsilon$ si $d(F, G) < \eta$. On admet alors H vérifiée.

– $d^2(G, Fn) \geqslant (q - 1)/(n\epsilon)$ avec une probabilité supérieure ou égale à $1 - \epsilon$ si $d(F, G) \geqslant \eta$. On admet alors que H n'est pas vérifié.

Dans le cas présent G est la distribution normale de moyenne μ et d'écart type σ. Si la réponse à la question posée plus haut est affirmative la valeur trouvée du paramètre k définira une transformation normalisatrice des données, dans le cas contraire aucune transformation de BOX et COX n'a le pouvoir de normaliser convenablement les données.

Transformations de BOX et COX à un seul paramètre : Lorsque, comme c'est souvent le cas en pratique, k2 est connu, à la translation initiale des données prés, qui à x associe x + k2, l'étude de la transformation de BOX et COX associée se ramène à celle d'une transformation du même type à un seul paramètre réel k. En prenant k1 = k et k2 = 0 dans les expressions précédentes on obtient alors les relations appropriées.

Les considérations qui suivent se situent dans le cadre des transformations de BOX et COX à un seul paramètre.

Précision et robustesse de l'estimateur de k : Même dans le cas favorable où il existe un paramètre normalisateur k, les valeurs prises par les données transformées appartiennent à un intervalle]A, B[qui, sauf lorsque la transformation est logarithmique, est une partie seulement de la droite réelle. La variable transformé Y ne peut donc suivre au mieux qu'une loi normale tronquée [30].

En fait DRAPER et COX [11] ont montré que la variance de l'estimateur du maximum de vraisemblance, $\hat{k}$ de k, s'exprime comme le produit de deux termes dont le second représente la part due à la non normalité des données transformées (il est égal à 1 si celles-ci sont effectivement normales).

$$V(\hat{k}) = (2/(3n\theta^2))(1 - (\gamma 1^2/3) + (7\gamma 2/18))^{-1}$$

où $\gamma 1 = b1$ et $\gamma 2 = b2 - 3$ représentent les coefficients d'asymétrie et d'aplatissement et $\theta = k\sigma/(1 + k\mu)$.

Il est à souligner, d'une part que les intervalles de confiance obtenus par le rapport du maximum de vraisemblance ne sont valables qu'asymptotiquement, d'autre part que l'estimateur $\hat{k}$ est sensible aux points aberrants.

Ces remarques ont poussé ANDREWS [1] à proposer l'utilisation d'un test de FISHER, moins sensible aux points aberrants, pour tester l'hypothèse $k = k_0$. Cette méthode est critiquée par ATKINSON [5] qui montre expérimentalement que pour une distribution normale le test du rapport de vraisemblance est plus puissant. CAROLL [10] considère que les transformations puissances conduisent à des distributions qui sont normales dans leur partie centrale, mais qui possèdent des extrêmes plutôt exponentielles. Ainsi si :

$$p(x) = \begin{cases} x^2/2 & \text{pour } |x| \leqslant h \\ h(|x| - (h/2)) & \text{pour } |x| > h \end{cases}$$

il obtient, en utilisant la méthode du maximum de vraisemblance avec la famille des distributions proportionnelles à $\exp(-\ p(x))$, une méthode plus robuste que la méthode du maximum de vraisemblance de BOX et COX et plus puissante, mais moins robuste, que celle proposée par ANDREWS.

Propriété particulière de l'estimateur du maximum de vraisemblance : Dans le cadre des modèles linéaires BOX et COX ont montré que la solution obtenue pour k par la méthode du maximum de vraisemblance apparait comme une sorte de compromis entre les différentes puissances suggérées (voir ANSCOMBE [4]) dans l'analyse des résidus par :

– l'hétérogénéité de la variance ;
– la non additivité ;
– la non normalité.

Transformations exponentielles [26]

Définition 1 : On appelle transformations exponentielles la famille des transformations à trois paramètres a, b, c réels définis pour tout x réel par :

$$x \longrightarrow a + b \,.\, \exp(cx)$$

Ces transformations constituent un cas particulier des transformations de Johnson.

Propriété : [12] La famille des transformations puissances génère comme cas limite la famille des transformations exponentielles.

En effet pour x supérieur à $-$ t on peut écrire si $t > 0$

$$r + s(x + t)^u = r + s(t(1 + x/t))^u = r + st^u((1 + x/t)^t)^{u/t}$$

Dans la pratique t est le paramètre de translation des données initiales qui permet de les rendre positives, c'est donc un réel positif. En posant $c = u/t$, $b = s.\exp(u.\log(t))$ et $a = r$ et en remarquant que pour x fixé supérieur à $-$ t, la limite lorsque t tend vers l'infini de $\exp(t.\log(1 + x/t))$ est égale à $\exp(x)$, il vient pour t 'très grand' :

$$r + s(x + t)^u \simeq a + b \,.\, \exp(cx)$$

Conséquences : Lorsque la variable initiale prend des valeurs négatives trés grandes les transformations exponentielles apparaissent comme cas limite des transformations puissances (k2 tend vers $-\infty$).

Une sous famille particulière est obtenue en prenant : $c = k$ et $b = -a = 1/c$ si k n'est pas nul. On note alors pour k non nul :

$$f_k(x) = \frac{\exp(kx) - 1}{k} = g_k(\exp(x))$$

Pour tout x fixé l'expression précédente tend vers x lorsque k tend vers zéro aussi est-il naturel de considérer les transformations suivantes.

Définition 2 : On appelle transformations exponentielles à un seul paramètre k réel la famille des transformations définies sur la droite réelle par :

$$f_k(x) = \begin{cases} x & \text{si } k = 0 \\ (\exp(kx) - 1)/k & \text{sinon} \end{cases}$$

Des propriétés analogues à celles mises en évidence pour les transformations de BOX et COX à un seul paramètre, existent pour les transformations exponentielles et entraînent des conséquences du même type. En particulier :

L'application qui au couple (x, k) associe le transformé de x par la transformation exponentielle de paramètre k est continue et dérivable en k. Ce qui assure pour tout x fixé une certaine stabilité par rapport à k.

La transformation exponentielle de paramètre k est dérivable sur la droite réelle de dérivée : exp(kx), elle est convexe (respectivement concave) si k est positif (respectivement négatif). Sa fonction inverse est le logarithme de l'inverse de la transformation puissance de paramètre k.

L'estimation de k par la méthode du maximum de vraisemblance conduit à l'expression suivante du même type que celle fournie par BOX et COX.

$$\mathrm{Lm}(k) = k \sum_{i=1}^{n} Xi - n \cdot \log(\hat{\sigma}_k)$$

Généralisation à la normalisation de variables vectorielles [16]

Il est possible de généraliser l'utilisation de transformations paramétriques de type Box et Cox ou exponentiel au cas de variables vectorielles X = (X1, X2, ... Xp). On recherche alors le vecteur k = (k1, k2, ..., kp) tel que les données obtenues après transformation par Tk = (Tk1 , Tk2 , ..., Tkp) puissent être considérées comme suivant une loi gaussienne. Des méthodes d'estimation de k par le maximum de vraisemblance découlent simplement de la généralisation de celles utilisées pour le cas unidimensionnel. Elles font intervenir l'estimateur de la matrice de variances-covariances des marginales transformées.

En fait comme l'indique GNANADESIKAN [16] il est souvent préférable d'essayer de normaliser les marginales. Théoriquement cette opération ne normalise pas forcément la loi conjointe, mais l'expérience montre que dans de nombreux cas il en est ainsi.

Retour aux données initiales après une transformation normalisatrice

Rendre les résultats d'une analyse interprétables nécessite souvent de les exprimer dans les mêmes unités que les données initiales. Dans les cas où l'expression des quantiles suffit à l'interprétation, le fait que pour une transformation croissante T on ait :

$$X_f = T^{-1}(Y_f)$$

permet d'obtenir aisément l'estimation des quantiles de la variable initiale en fonction de ceux de la variable transformée.

Cependant l'interprétation des résultats nécessite en général les valeurs de la moyenne et de la variance de la variable initiale. La variable transformée étant supposée normalement distribuée les estimateurs :

$$\bar{Y} \quad \text{et} \quad S_Y^2$$

sont des estimateurs efficaces de E(Y) et V(Y). Comme l'espérance de X est aussi l'espérance de l'image de Y par la transformation inverse de T, un estimateur efficace de $E(X)$ sera forcément fonction de ces deux estimateurs. Il est ainsi logique de penser en premier lieu à prendre comme estimateur de $E(X)$:

$$T^{-1}(E(Y))$$

Cependant en général :

$$E(T^{-1}(Y)) \quad \text{est différent de } T^{-1}(E(Y))$$

Ainsi prendre pour estimateur (respectivement intervalle de confiance $I\alpha$) de $E(X)$, l'image inverse de l'estimateur (respectivement de l'intervalle de confiance $J\alpha$) de $E(Y)$, présente comme danger que rien ne garantit qu'il existe une valeur appartenant à $J\alpha$ dont l'image par la transformation inverse soit égale à l'espérance de X.

Pour pallier cet inconvénient PATTERSON [29] propose alors de trouver une fonction h telle que l'image de $E(Y) + h(Y)$ par la transformation inverse T^{-1} soit à peu près égale à $E(X)$. L'expression de h est simple à trouver dans le cas de la transformation logarithmique : $h(Y) = V(Y)/2$ mais PATTERSON ne donne pas de méthode générale pour déterminer h.

LAND [24] donne un bref aperçu des différentes techniques existant dans la littérature statistique qui permettent d'obtenir une estimation de l'intervalle de confiance de la moyenne de X, HOYLE [19] développe en plus une méthode d'estimation de la variance. Ces résultats s'obtiennent en utilisant le développement en série entière, au voisinage de $\mu = E(Y)$, de l'image de Y par l'inverse de T.

Exemple d'utilisation

Pour illustrer l'efficacité des procédés de normalisation exposés, la méthode de BOX et COX a été appliquée aux relevés de durées 4 et 6 heures de la concentration en anhydride sulfureux, sur une station proche du site de Lacq, pour l'année 1977. Pour chacune de ces durées d'observation on a représenté, afin de visualiser l'effet normalisateur, l'histogramme des données brutes puis celui des données transformées : figures 1 à 4 de l'annexe 2.

CONCLUSION

Le présent article ne constitue pas une liste exhaustive de toutes les techniques permettant de se ramener à un modèle normal. D'autres méthodes existent par ailleurs notamment celles basées sur la recherche de codages normaux (22), souvent utilisées pour l'étude des dépendances entre variables qualitatives. Cependant l'ensemble des procédés exposés ici représente un éventail des principales démarches utilisées dans le domaine de la normalisation des données. Atteindre la normalité ne doit, de toute façon, pas constituer une fin en soi, mais un moyen de poursuivre une analyse, ni devenir une obsession.

ANNEXES

Annexe 1

• Tableau-I-

variance en fonction de moyenne	transformation	variance de la variable transformée	loi sous-jacente
m	$\sqrt{t}$	.25	Poisson *
$k^2 m$		$.25\ k^2$	
$2m^2/(n-1)$	Log(t)	$2/(n-1)$	
$k^2\ m^2$	Log(t) ou Log(t+1)	k^2	
$m(1-m)/n$	arcsin $\sqrt{t}$ rd	.25/n	binomiale **
	arcsin $\sqrt{t}$ dg	821/n	
$m(1-m)/n$	• probit • Log(t/(1-t)) (logit)	pas cste $1/(mn(1-m))$	binomiale
$k^2\ (1-m)^2\ m^2$	logit	k^2	
$(1-m^2)^2/(n-1)$	$\frac{1}{2}\log\left(\frac{1+t}{1-t}\right)$	$1/(n-3)$	
$m+k^2\ m^2$	(1/k).argsh(k$\sqrt{t}$)	.25	binomiale négative ***
$h^2\ (m+k^2\ m^2)$		$.25\ h^2$	

Remarques : ANSCOMBE [3] a montré que :

Si t suit une loi de Poisson la transformation $\sqrt{t + (3/8)}$ permet de mieux stabiliser la variance

** Si t suit une loi binomiale une meilleure stabilisation est obtenue avec

$$\text{Arcsin}\left\{\left(\frac{nt + (3/8)}{n + (3/4)}\right)^{1/2}\right\}$$

*** Si t suit une loi binomiale négative une meilleure transformation stabilisatrice est :

$$(1/k)\ .\ \text{Argsh}\left\{k\left(\frac{nt + (3/8)}{n - (3/4)}\right)^{1/2}\right\}$$

Annexe 2

Histogrammes pour la durée 4 heures

Le nombre de données retenues est égal à 484

classe de	classe a	nbre	%	
2.50	90.75	382	78.93	***************************************
90.75	179.00	51	10.54	*****
179.00	267.25	17	3.51	**
267.25	355.50	11	2.27	*
355.5	443.75	12	2.48	*
443.75	532.00	5	1.03	*
532.00	620.25	1	0.21	
620.25	708.50	2	0.41	
708.50	796.75	1	0.21	
796.75	885.00	2	0.41	

Minimum = 2.50
Maximum = 885.00
Moyenne = 69.71
Ecart-type = 120.94
b1 = 11.05
b2 = 16.37

Figure 1. – Histogramme des données brutes.

classe de	classe a	nbre	%	
0.876	1.281	42	8.68	****
1.281	1.686	46	9.50	*****
1.686	2.091	59	12.19	******
2.091	2.496	81	16.74	********
2.496	2.901	59	12.19	******
2.901	3.306	56	11.57	******
3.306	3.711	53	10.95	*****
3.711	4.116	39	8.06	****
4.116	4.521	35	7.23	****
4.521	4.926	14	2.89	*

Minimum = 0.88
Maximum = 4.93
Moyenne = 2.69
Ecart-type = 1.02
b1 = 0.74E-03
b2 = 2.24

Figure 2. – Histogramme des données transformées : la transformation de Box et Cox a pour paramètre – 0.1.

Histogrammes pour la durée 6 heures

Le nombre de données retenues est égal à 383

classe de	classe a	nbre	%	
1.67	75.50	304	79.37	**
75.50	149.33	39	10.18	*****
149.33	223.17	16	4.18	**
223.17	297.00	7	1.83	*
297.00	370.83	7	1.83	*
370.83	444.67	3	0.78	
444.67	518.50	3	0.78	
518.50	592.33	0	0.00	
592.33	666.17	3	0.78	
666.17	740.00	1	0.26	

Minimum = 1.67
Maximum = 740.00
Moyenne = 58.60
Ecart-type = 101.91
b1 = 11.49
b2 = 16.65

Figure 3. – Histogramme des données brutes.

classe de	classe a	nbre	%	
0.502	0.981	27	7.06	****
o.981	1.459	30	7.83	****
1.459	1.938	43	11.23	******
1.938	2.417	60	15.67	********
2.417	2.896	60	15.67	********
2.896	3.374	53	13.84	*******
3.374	3.853	43	11.23	******
3.853	4.332	34	8.88	****
4.332	4.811	23	6.01	***
4.811	5.289	10	2.61	*

Minimum = 0.50
Maximum = 5.29
Moyenne = 2.69
Ecart-type = 1.17
b1 = 0.13E-03
b2 = 2.31

Figure 4. – Histogramme des données transformées : la transformation de Box et Cox a pour paramètre – 0.07.

BIBLIOGRAPHIE

[1]ANDREWS. – A note on the selection of data transformations. (1971) Biometrika num. 58, p. 249-254.

[2]ANDREWS, GNANADESIKAN and WARNER. – Transformation of multivariate data (1971). *Biometrics num. 27,* p. 825-840.

[3]ANSCOMBE. – The transformation of Poisson, binomial and negative binomial data. (1948). *Biometrika num. 35,* p. 246.

[4]ANSCOMBE & TUKEY. – Examination and analysis of residuals (1963). *Technometrics num. 5*, p. 141-160.

[5]ATKINSON. – Testing transformations to normality (1973). *Journal of the Royal Statistical Society,* vol. B-35, p. 473-479.

[6]BARTLETT. – The use of transformations (1947), *Biometrics num. 1,* p. 39-52.

[7]BICKEL & DOKSUM. – An analysis of transformation revisited (1981). *JASA,* vol. 76, num. 374, p. 296-311.

[8]BLOM. – Transformation of the binomial, negative-binomial, Poisson and khi2 distributions. (1954) *Biometrika num. 41.*

[9]BOX & COX. – An analysis of transformations (1964). *Journal of the Royal Statistical Society,* Vol. B-26, p. 211-252.

[10]CAROLL. – A robust method for testing transformation to achieve approximate normality (1980). *Journal of the Royal Statistical Society,* vol: B-42, num. 1, p. 71-78.

[11]COX & DRAPER. – On distributions and their transformation to normality (1969). *Journal of the Royal Statistical Society,* vol. B-31, p. 472-476.

[12]CRESSIE. – The exponential and power data transformations. *The Statistician,* vol. 27, num. 1, p. 57-60.

[13]CURTISS. – On transformations used in the analysis of variance (1943). *Annals of Mathematical Statistics,* num. 14.

[14]DRAPER & HUNTER. – Transformations some examples revisited (1969). *Technometrics,* num. 11, p. 23-40.

[15]FRIEDMAN. – The use of ranks to avoid the assumption of normality implicit in the analysis of variance (1937). *JASA,* vol. 32.

[16]GNANADESIKAN. – *Method for statistical data analysis of multivariate observations* (1977). Wiley & sons.

[17]HALDANE – The approximate normalisation of a class of frequency distribution (1938). *Biometrika,* num. 29.

[18]HINKLEY. – On power transformation to symmetry. (1975) *Biometrika,* vol. 62, p. 101-111.

[19]HOYLE. – The estimation of variances after using a gaussianating transformation (1968). *Annals of Mathematical Statistics,* vol: 39, num. 4, p. 1125-1143.

[20]JOHNSON. – Tables to facilitate fitting Su frequency curves (1965). *Biometrika,* num. 52, p. 547-558.

[21]JOHNSON & KOTZ. – Continuous univariate distribution (1970). *Houghton-Mifflin*, vol. II & III.

[22]KENDALL & STUART. – *The advanced theory of statistics.* Griffin-Books London vol. I, II & III.

[23]KIMBER. – Tests for a single outlier in a gamma sample with unknown shape and scale parameters (1979). *Applied Statistics*, p. 243-250.

[24]LAND. – Confidence interval estimation for means after data transformations to normality (1974). *JASA*, vol : 69, num. 347, p. 795-802.

[25]LINDSEY. – The roles of transformations to normality (1975). *Biometrics*, num. 31, p. 247-249.

[26]MANLY. – Exponential data transformations (1976). *The statistician*, vol. 25, num. 1, p. 37-42.

[27]MATUSITA. – Decision rules, based on the distance, for problem of fit two samples and estimation (1955). *Annals of the Institute of Statistical Mathematics*, p. 631-640.

[28]NEYMAN & SCOTT. – Corrections for biais introduced by a transformation of variables (1960). *Annals of Mathematical Statistics*, vol. 31, p. 643-655.

[29]PATTERSON. – Difficulties involved in the estimation of a data population mean using transformed sample data (1966). *Technometrics*, vol. 8, num. 3, p. 535-537.

[30]POIRIER. – The use of the Box and Cox transformation in limited dependent variable models (1978). *JASA*, vol. 73, num. 362, p. 284-287.

[31]SCHLESSELMAN. – Power families: a note on the Box and Cox transformation (1961). *Journal of the Royal Statistical Society*, vol. B-33.

[32]WHITMORE & YALOVSKY. – A normalizing logarithmic transformation for inverse gaussian random variables (1978). *Technometrics*, vol. 20, num. 2, p. 207-208.

[33]WILSON & HILFERTY. – *The distribution of Chi square* (1931). Proceedings of National Academy of Sciences-USA num. 17, p. 684-688.

[34]WINTERBON. – Determining parameters of the Johnson Su distributions (1978). *Commun. Stat. Simula. Computa.*, vol. B7, num. 3, p. 223-226.

[35]ZAREMBKA. – Transformations of variables in econometrics. Frontiers in Econometrics, Ed. P-Zarembka – New York Academic Press.

REMERCIEMENTS

Le travail exposé ici a été en partie réalisé dans le cadre d'une collaboration du laboratoire de statistique du Centre de Recherche en Informatique et Gestion de Montpellier et la Société Nationale Elf Aquitaine. Certains aspects ont fourni l'idée centrale d'une thèse de troisième cycle de l'auteur consacrée à l'élaboration d'un modèle de pollution atmosphérique. En conséquence nous tenons à exprimer nos sincères remerciements à Messieurs BAPSERES et CAMPS de la SNEA pour avoir rendu possible ce travail et à adresser notre chaleureuse reconnaissance au Professeur Yves ESCOUFIER pour l'attention amicale qu'il a bien voulu accorder à nos travaux.