

NOTES DE STATISTIQUE ET D'INFORMATIQUE

2011/3

MACROS MINITAB POUR LE CHOIX D'UNE
TRANSFORMATION POUR LA
NORMALISATION DE VARIABLES

R. PALM, Y. BROSTAUX, J. J. CLAUSTRIAUX

Université de Liège – Gembloux Agro-Bio Tech

Unité de Statistique, Informatique et Mathématique

appliquées à la bioingénierie

GEMBLOUX

(Belgique)

MACROS MINITAB POUR LE CHOIX D'UNE TRANSFORMATION POUR LA NORMALISATION DE VARIABLES

R. PALM^{*}, Y. BROSTAUX[†] et J. J. CLAUSTRIAUX[‡]

RÉSUMÉ

Différentes macros sont proposées aux utilisateurs de Minitab pour la recherche d'une transformation permettant de normaliser une variable.

Les transformations étudiées sont la transformation logarithmique, les transformations de BOX et COX avec les valeurs λ égales à -2 , -1 , $-0,5$, $0,5$ et 2 , la transformation de BOX et COX avec la valeur de λ estimée par le maximum de vraisemblance et les transformations de JOHNSON.

L'utilisation des macros est illustrée par deux exemples numériques.

SUMMARY

Macros are proposed to Minitab users for transforming data to normality.

The following transformations are examined : log transformation, BOX et COX transformations with λ equal to -2 , -1 , -0.5 , 0.5 and 2 , BOX and COX transformation with maximum likelihood estimation of λ and transformations using the JOHNSON system of distributions.

Two examples of using the macros are given.

1. INTRODUCTION

De nombreuses méthodes statistiques reposent sur l'hypothèse de normalité des distributions des variables étudiées, mais en pratique, celles-ci ne répondent pas toujours à cette condition.

^{*}Professeur à l'Université de Liège, Gembloux Agro-Bio Tech.

[†]Chef de Travaux et Chargé de cours à l'Université de Liège, Gembloux Agro-Bio Tech.

[‡]Professeur ordinaire à l'Université de Liège, Gembloux Agro-Bio Tech.

Dans de telles situations, trois possibilités s'offrent au praticien. La première solution consiste à utiliser les méthodes en dépit de cette non-normalité, en considérant que les résultats obtenus sont approchés.

La deuxième solution est de remplacer une méthode d'analyse reposant sur l'hypothèse de normalité par une autre méthode, pour laquelle la normalité n'est pas une condition d'application.

Enfin, la troisième solution repose sur la transformation des variables non normales en variables normales ou approximativement normales.

Le choix de la solution dépend de plusieurs facteurs, parmi lesquels on peut citer :

- la robustesse au non-respect de la normalité de la méthode statistique considérée,
- la disponibilité de méthodes alternatives,
- l'objectif poursuivi par l'étude,
- la connaissance *a priori*, par expérience ou sur une base théorique, d'une transformation permettant de normaliser les données,
- le nombre d'observations disponibles.

Lorsque l'utilisateur a porté son choix sur la transformation de variable afin de résoudre le problème de la non-normalité mais qu'il n'a pas une connaissance *a priori* de la nature de la transformation à utiliser, il procède généralement par essais et erreurs afin de trouver une transformation satisfaisante. Une telle recherche peut être assez longue et fastidieuse.

Afin d'automatiser cette recherche, l'Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie de GxABT propose aux utilisateurs du logiciel Minitab des macros permettant d'analyser diverses transformations. Ces macros, ainsi que les notices d'utilisation, sont disponibles sur le site web de l'Unité :

<http://www.fsagx.ac.be/si/>

en cliquant sur le lien Macros, puis sur le thème en question.

Nous présentons d'abord les deux familles de transformations considérées : les transformations de BOX et COX, qui englobent, comme cas particuliers, les transformations classiques (paragraphe 2) et les transformations de JOHNSON (paragraphe 3). Nous décrivons ensuite deux macros analysant diverses transformations (paragraphe 4). Enfin, nous tirons quelques conclusions (paragraphe 5).

Pour illustrer les transformations et l'utilisation des macros, nous considérons deux exemples. Le premier est repris de DAGNELIE [2006]. Il concerne le poids, en grammes, de 99 poissons d'une espèce donnée, observés dans un biotope particulier. Le deuxième exemple provient de PEARN *et al.* [2004]. Il concerne l'étude de l'aptitude d'un procédé de production de composants électroniques. La variable d'intérêt est une caractéristique technique d'amplificateurs, mesurée

en décibels. On dispose de 120 mesures et on s'intéresse à trois valeurs particulières, qui représentent respectivement la tolérance inférieure ($t_i = 7,75$), la moyenne cible ($m_c = 10$) et la tolérance supérieure ($t_s = 12,25$) du procédé. Cet exemple a déjà été utilisé antérieurement par CHOU *et al.* [1998], pour illustrer l'utilisation des transformations de JOHNSON.

On notera aussi que le logiciel Minitab propose des commandes ou des options dans des commandes réalisant les transformations de BOX et COX et les transformations de JOHNSON, en relation avec les outils utilisés pour le contrôle de la qualité. Ces commandes ne permettent cependant pas une comparaison directe des différentes transformations, comme le font les macros présentées au paragraphe 4.

Les macros proposées considèrent que la colonne des observations correspond à un seul échantillon et qu'il n'y a pas de regroupements logiques dans les données. Ces macros ne devraient donc pas être utilisées si la colonne est constituée d'un empilement de plusieurs échantillons, comme cela serait le cas pour des données devant être soumises à une analyse de la variance. Dans une telle situation, une solution serait par exemple de soumettre à l'analyse les résidus d'un modèle d'analyse de la variance, ceux-ci ne comportant plus d'effets liés aux facteurs. Il faudrait cependant être attentif au fait que la transformation de variable modifie la variabilité. Il serait par conséquent utile de vérifier l'homoscédasticité après transformation de la variable.

De même, les macros ne sont pas utilisables dans le contexte de la régression, la transformation des résidus en vue de leur normalisation ne donne en effet pas la transformation à appliquer à la variable à expliquer et aux variables explicatives. La transformation de BOX et COX de la variable à expliquer est cependant une technique classiquement proposée, mais l'estimation du paramètre intervenant dans la transformation se fait de manière différente de celle présentée dans cette note. Des informations à ce sujet sont données notamment par DRAPER et SMITH [1998], PALM et IEMMA [2002] et WEISBERG [2005]. Une macro Minitab permettant de rechercher la transformation dans le cas de la régression est disponible sur le site web de l'unité, à l'adresse donnée ci-dessus [PALM, 2002].

2. TRANSFORMATION DE BOX ET COX

2.1. Principe

La transformation de Box et Cox [1964] s'écrit :

$$z(\lambda) = \begin{cases} x^\lambda & \text{si } \lambda \neq 0 \\ \log_e x & \text{si } \lambda = 0. \end{cases}$$

En donnant à λ respectivement les valeurs 0,5, 0 et -1 , on retrouve les transformations racine carrée, logarithmique et inverse, qui sont des transformations assez classiques.

En l'absence d'une base théorique permettant de sélectionner une valeur particulière de λ , on peut déterminer λ à partir des observations elles-mêmes, par la méthode du maximum de vraisemblance.

Dans ce but, on écrit la transformation de BOX et COX sous une forme un peu plus compliquée :

$$w(\lambda) = \begin{cases} (x^\lambda - 1) / \lambda \bar{x}_g^{\lambda-1} & \text{si } \lambda \neq 0 \\ \bar{x}_g \log_e x & \text{si } \lambda = 0. \end{cases}$$

Dans cette expression, $\bar{x}_g$ est la moyenne géométrique de x . Cette seconde présentation a l'avantage d'être continue en fonction de λ et de permettre la comparaison directe des variances des données transformées par différentes valeurs de λ .

Le logarithme de la fonction de vraisemblance s'écrit alors :

$$\log_e L(\lambda) = -\frac{n}{2} \log_e SCE(\lambda),$$

avec :

$$SCE(\lambda) = \sum_{i=1}^n [w_i(\lambda) - \bar{w}(\lambda)]^2.$$

Estimer λ au sens maximum de vraisemblance revient à trouver la valeur de λ qui maximise la fonction de vraisemblance ou son logarithme ou encore qui minimise la variance de $w(\lambda)$.

En pratique, on fait varier λ , généralement de -2 à 2 , et, pour chaque valeur de λ , on détermine la variance de $w(\lambda)$. On retient alors la valeur de λ pour laquelle la variance est minimum.

Il est possible aussi de déterminer, de manière approximative, l'intervalle de confiance de λ . Cet intervalle comprend toutes les valeurs de λ telles que :

$$\log_e L(\lambda) > \log_e L(\hat{\lambda}) - \frac{1}{2} \chi_{1-\alpha}^2,$$

$L(\hat{\lambda})$ étant la valeur maximum de la fonction de vraisemblance et $\chi_{1-\alpha}^2$ étant le pourcentile $1 - \alpha$ de la distribution χ^2 à un degré de liberté. Les limites, λ_1 et λ_2 , sont déterminées à partir du graphique donnant la vraisemblance en fonction de λ ou par interpolations linéaires à partir des couples $L(\lambda)$ et λ . Une fois la valeur de λ déterminée, la transformation de la variable se fera, en pratique, en utilisant la transformation présentée en premier lieu.

Bien que la transformation de BOX et COX permette la normalisation de beaucoup de distributions observées, l'utilisateur n'a pas la garantie que toute série peut être normalisée par cette transformation.

2.2. La macro BOXCOXTRANS

La macro BOXCOXTRANS permet la recherche de la valeur de λ par la méthode du maximum de vraisemblance.

Cette estimation se fait en deux temps. La vraisemblance est d'abord calculée pour les valeurs de λ variant de λ_i à λ_s avec un pas de variation égal à δ . Soit $\hat{\lambda}_0$ la valeur qui, parmi les valeurs testées, maximise la vraisemblance. Une nouvelle recherche est ensuite réalisée en faisant varier λ de $\hat{\lambda}_0 - \delta$ à $\hat{\lambda}_0 + \delta$, avec un pas de variation égal à $\delta/10$. La valeur finalement retenue pour $\hat{\lambda}$ est celle qui maximise la vraisemblance pour ce nouvel ensemble de valeurs testées.

La macro donne, sous forme graphique, l'évolution en fonction de λ de la vraisemblance et de la corrélation entre les données transformées et les scores normaux correspondants. Elle réalise également la transformation des données pour la valeur $\hat{\lambda}$ et permet aussi de transformer, pour ce même $\hat{\lambda}$, d'éventuelles données supplémentaires. Enfin, différentes options permettent de contrôler le volume des résultats affichés.

Des informations complémentaires sont données dans la notice d'utilisation qui accompagne la macro.

La macro a été appliquée à la série des poids des poissons présentée dans l'introduction et la figure 1 reprend les informations concernant la transformation optimale. Les calculs ont été réalisés en faisant varier λ de -2 à 2 , et en donnant à δ la valeur $0,1$ (valeur par défaut).

Le maximum du logarithme de la vraisemblance est atteint pour $\hat{\lambda} = 0,37$. En pratique, on pourrait donc réaliser la transformation :

$$z = x^{0,37}.$$

Pour cette transformation, la corrélation de z et des scores normaux est égale à $0,9882$.

La valeur théorique du test de normalité de RYAN-JOINER pour 99 observations et pour un niveau de signification de 5 % est égale à $0,9870$. On ne rejette donc pas l'hypothèse de normalité de la variable transformée, alors que l'hypothèse de normalité est rejetée pour la variable non transformée, la corrélation avec les scores normaux étant dans ce cas égale à $0,9334$, comme nous le verrons au paragraphe 4.1.

Pour $\lambda = 0,37$, le logarithme de la fonction de vraisemblance $\log_e L(\lambda)$ est égal à $-396,27$ et l'intervalle de confiance pour un degré de confiance de $0,95$ comprend les valeurs de λ pour lesquelles le logarithme de la vraisemblance est supérieur à :

$$-396,27 - \left(\frac{1}{2}\right) 3,84 = -398,19,$$

```
%BOXCOXTRANS C1
```

Valeurs théoriques des corrélations entre observations et scores normaux :

```
Nobs = 99 cor10 = 0.9893 cor05 = 0.9870 cor01 = 0.9816
```

```
Lambda optimal= 0.37 (Lim. Conf.= 0.18 et 0.58)  
LOG Vrais.= -396.27 Corr. scores norm.= 0.9882
```

Figure 1 – Poids des poissons : commande et résultats de la macro BOXCOX-TRANS.

soit :

$$\lambda_1 = 0,18 \quad \text{et} \quad \lambda_2 = 0,58$$

Ces limites de confiance englobent la valeur $\lambda = 0,5$. Si on souhaitait une transformation plus classique, on pourrait retenir cette dernière valeur et la transformation s'écrirait :

$$z = \sqrt{x}.$$

Pour cette transformation, la corrélation avec les scores normaux est égale à 0,9855, soit une valeur très légèrement inférieure à la valeur théorique du test de RYAN-JOINER, comme nous le verrons au paragraphe 4.

Les figures 2 et 3 donnent l'évolution du logarithme de la vraisemblance et de la corrélation entre les données transformées et les scores normaux, en fonction de λ .

3. TRANSFORMATIONS DE JOHNSON

3.1. Principe

Les transformations de JOHNSON [1949] sont nettement plus complexes que les transformations de BOX et COX, dans la mesure où elles dépendent, non plus d'un seul paramètre, mais de quatre paramètres. De plus, elles reposent sur trois fonctions particulières, qui visent à transformer les observations de départ x en valeurs provenant d'une variable normale réduite.

Ces trois fonctions constituent les trois familles du système de courbes de JOHNSON, qui sont notées S_B , S_L et S_U . Ces fonctions, ainsi que les contraintes sur les paramètres et sur x sont reprises dans le tableau 1.

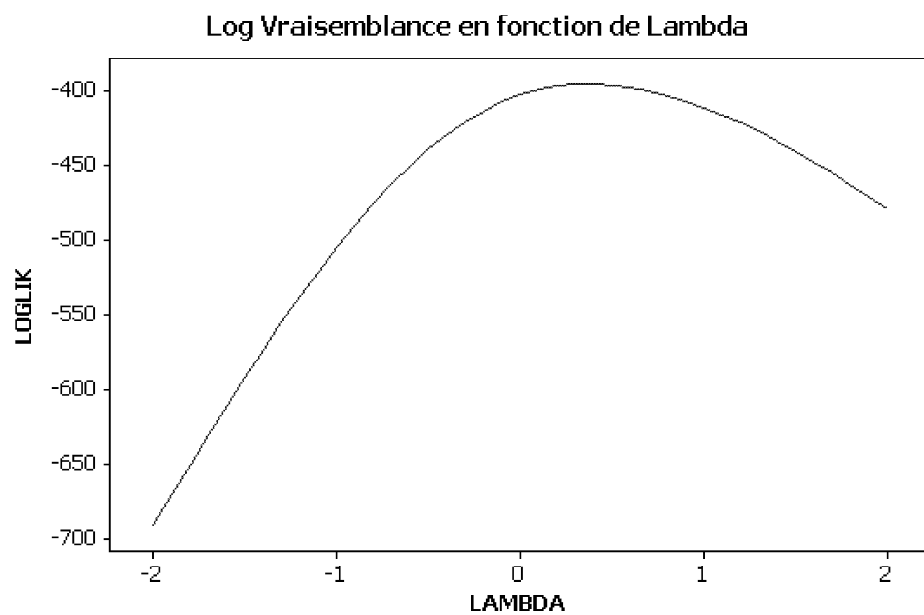


Figure 2 – Poids des poissons : évolution du logarithme de la vraisemblance en fonction de λ .

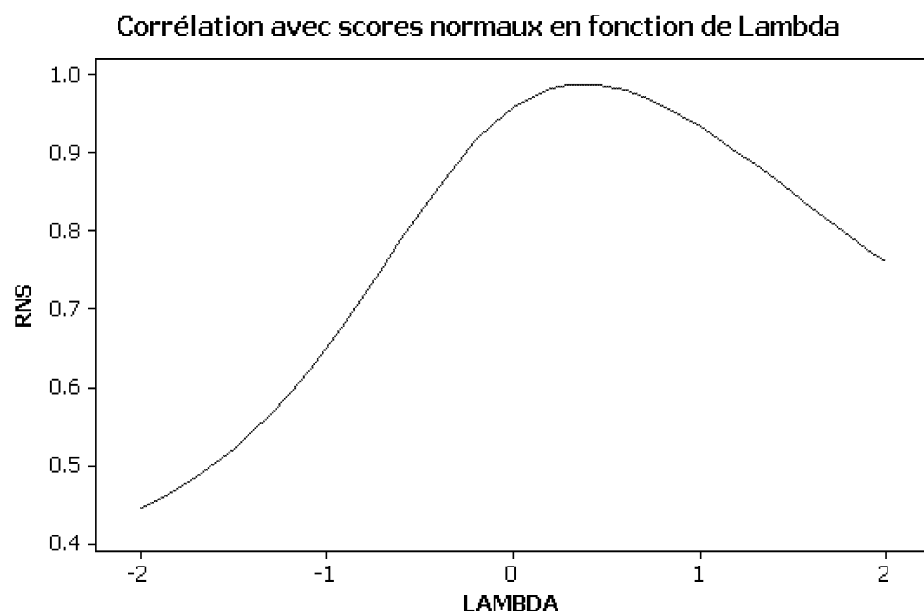


Figure 3 – Poids des poissons : évolution de la corrélation entre les données transformées et les scores normaux, en fonction de λ .

On peut constater que dans le cas de la transformation S_B , les valeurs de x sont bornées¹, alors qu'elles ne le sont pas pour la transformation S_U ². Pour la transformation S_L , les valeurs de x sont bornées inférieurement et sont de type lognormal³.

Tableau 1 – Transformations de JOHNSON.

Famille	Transformation	Conditions sur les paramètres	Conditions sur x
S_B	$z = \gamma + \eta \log_e \left(\frac{x-\varepsilon}{\lambda+\varepsilon-x} \right)$	$\eta, \lambda > 0,$ $-\infty < \gamma, \epsilon < \infty,$	$\varepsilon < x < \varepsilon + \lambda$
S_L	$z = \gamma + \eta \log_e (x - \varepsilon)$	$\eta > 0,$ $-\infty < \gamma, \epsilon < \infty,$	$x > \varepsilon$
S_U	$z = \gamma + \eta \sinh^{-1} \left(\frac{x-\varepsilon}{\lambda} \right)$	$\eta, \lambda > 0,$ $-\infty < \gamma, \epsilon < \infty,$	$-\infty < x < \infty$

On peut noter que la fonction S_L ne dépend que de trois paramètres, contrairement aux deux autres, qui font intervenir quatre paramètres.

Pour normaliser une série d'observations x_i par les fonctions de JOHNSON, l'utilisateur doit choisir parmi les trois fonctions celle qui est la plus adéquate et il doit ensuite en estimer les paramètres.

Différentes solutions ont été proposées et nous retenons la démarche décrite par CHOU *et al.* [1998], qui repose sur la détermination de quatre quantiles estimés à partir des observations et que nous notons $\hat{x}(p_i)$ ($i = 1, \dots, 4$).

Les probabilités p_i associées à ces quantiles correspondent aux quatre valeurs suivantes de la fonction de répartition de la distribution normale réduite :

$$p_1 = \Phi(-3u_0), p_2 = \Phi(-u_0), p_3 = \Phi(u_0) \quad \text{et} \quad p_4 = \Phi(3u_0),$$

u_0 étant un nombre fixé. Nous reviendrons sur le choix du u_0 ultérieurement.

A partir des quantiles estimés, on détermine le rapport des quantiles⁴ :

$$QR = \frac{[\hat{x}(p_4) - \hat{x}(p_3)][\hat{x}(p_2) - \hat{x}(p_1)]}{[\hat{x}(p_3) - \hat{x}(p_2)]^2}$$

¹En anglais : *bounded*.

²En anglais : *unbounded*.

³En anglais : *lognormal*.

⁴En anglais : *quantile ratio*.

qui permet d'orienter le choix de la famille de distributions à considérer. Il a en effet été démontré que si une variable aléatoire X possède une distribution de type S_B , alors le rapport théorique des quantiles est inférieur à 1. Il est égal à l'unité si X possède une distribution de type S_L et il est supérieur à l'unité si X possède une distribution de type S_U .

En pratique, on ne dispose cependant que du rapport des quantiles estimés. CHOU *et al.* [1998] suggèrent d'ajuster à la fois la distribution S_B et la distribution S_L si $\hat{QR} < 1$ et à la fois la distribution S_L et la distribution S_U si $\hat{QR} \geq 1$.

Pour chaque type de distribution, des formules permettent d'estimer les paramètres $\gamma, \eta, \varepsilon$ et λ à partir des quatre quantiles estimés $\hat{x}(p_i)$ [CHOU *et al.*, 1998].

Le choix du type de transformation et l'estimation des paramètres de la transformation dépendent donc des quatre quantiles estimés, qui, eux-mêmes, dépendent de la valeur de u_0 , définissant les probabilités p_i . CHOU *et al.* [1998] proposent de réaliser les calculs en faisant varier u_0 de 0,25 à 1,25, par pas de 0,01. Pour chaque valeur u_0 , on vérifie si, pour l'ajustement S_L toutes les observations sont supérieures à la valeur estimée du paramètre ε . Si ce n'est pas le cas, l'ajustement est éliminé. De même, si l'ajustement S_B a été réalisé, on vérifie si toutes les observations sont comprises entre les valeurs estimées de ε et de $\varepsilon + \lambda$ et, dans la négative, on élimine l'ajustement.

Pour les transformations qui respectent les conditions sur les x_i , on détermine alors un critère de qualité de la transformation en calculant, sur les données transformées, la statistique associée à un test de normalité et on retient la transformation pour laquelle la probabilité associée à la statistique est la plus grande. Ainsi, CHOU *et al.* [1998] proposent par exemple de retenir la transformation qui maximise la statistique W du test de SHAPIRO et WILK.

3.2. La macro JOHNSONTRANS

Cette macro recherche la transformation de JOHNSON selon la procédure décrite par CHOU *et al.* [1998], en remplaçant toutefois le test de normalité de SHAPIRO et WILK par le test de RYAN-JOINER.

La macro donne également, sous forme graphique, l'évolution en fonction de u_0 de la corrélation entre les données transformées et les scores normaux correspondants. Elle réalise aussi la transformation des données pour la transformation sélectionnée et permet de transformer, de la même manière, d'éventuelles données supplémentaires.

Des informations complémentaires sont données dans la notice qui accompagne la macro.

La figure 4 donne les résultats de la transformation des données relatives aux amplificateurs en considérant les trois données supplémentaires, reprises dans la colonne C11, qui correspondent à la tolérance inférieure, à la valeur cible et à la tolérance supérieure du procédé.

```
%JOHNSONTRANS c10;
XSUPPL c11.
```

Données supplémentaires

Row	Xs	ZS
1	7.75	-2.13563
2	10.00	1.01589
3	12.25	3.32048

Valeurs théoriques des corrélations entre observations et scores normaux :

```
Nobs = 120 cor10 = 0.9910 cor05 = 0.9891 cor01 = 0.9846
```

Transformation optimale

```
Type = Sb u0=0.65 Corr. scores norm.= 0.9989
```

```
ETA = 1.19718 GAMMA = 1.23997 LAMBDA = 5.66652 EPSILON = 7.43113
```

Figure 4 – Amplificateurs : commandes et résultats de la macro JOHNSON-TRANS.

La transformation retenue est la transformation S_B :

$$z = 1,23997 + 1,19718 \log_e \left(\frac{x - 7,43113}{5,66652 + 7,43113 - x} \right)$$

et, pour ces données transformées, la tolérance inférieure, la valeur cible et la tolérance supérieure sont respectivement égales à :

$$-2,13563, \quad 1,01589 \quad \text{et} \quad 3,32048.$$

La corrélation des données transformées avec les scores normaux est égale à 0,9989, soit une valeur supérieure à la valeur théorique, égale, pour un niveau de signification de 0,05, à 0,9891. On ne rejette donc pas l'hypothèse de normalité.

La figure 5 donne l'évolution de la corrélation des valeurs transformées avec les scores normaux en fonction de u_0 . On peut constater que la variation de la corrélation est très importante pour des valeurs de u_0 assez proches et que plusieurs valeurs de u_0 donnent des corrélations fort proches du maximum.

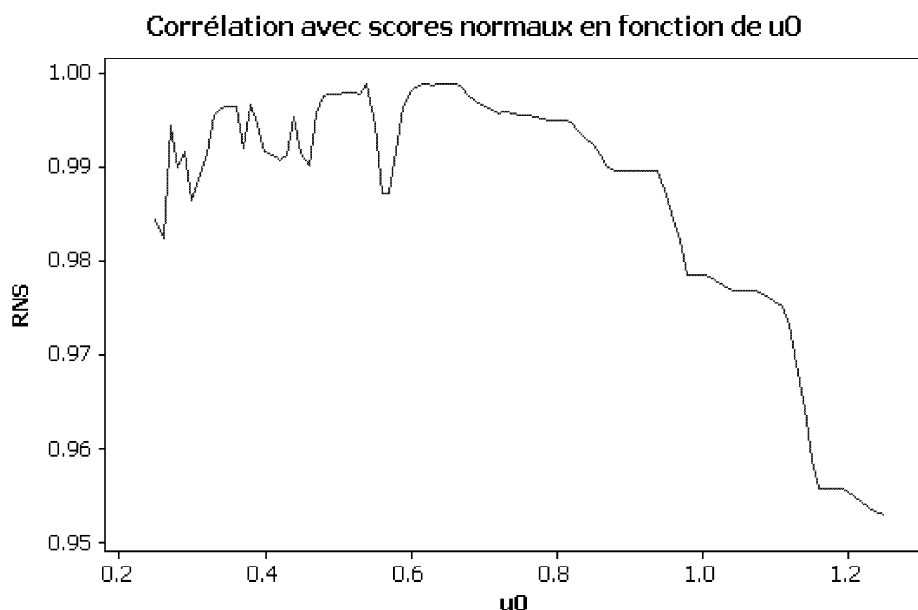


Figure 5 – Amplificateurs : évolution de la corrélation entre les données transformées et les scores normaux, en fonction de u_0 .

4. RECHERCHE D'UNE TRANSFORMATION

4.1. La macro ALLVARTRANSFO

La macro réalise de manière systématique diverses transformations de variables et calcule, pour chacune de celles-ci, la corrélation entre les données transformées et les scores normaux correspondants. Les six séries suivantes sont étudiées :

1. série initiale sans transformation ;
2. transformation logarithmique, en base e ;
3. transformation de BOX et COX avec une valeur de λ égale à -1, 0,5, 0 ou 0,5, la valeur retenue étant celle qui, parmi les quatre valeurs testées, maximise la vraisemblance ;
4. transformation de BOX et COX avec une valeur de λ égale à -2 ou 2, la valeur retenue étant celle qui maximise la vraisemblance ;
5. transformation de BOX et COX, avec recherche de la valeur de λ qui maximise la vraisemblance ;
6. transformation de JOHNSON.

Les données transformées peuvent être enregistrées dans le fichier. D'éventuelles données supplémentaires peuvent également être transformées en utilisant les paramètres des transformations définis à partir de la variable initiale.

Les corrélations entre les données transformées et les scores normaux peuvent être enregistrées, à la fois pour la variable initiale et pour les données supplémentaires. Pour la transformation de BOX et COX et pour la transformation de JOHNSON, les paramètres peuvent également être enregistrés.

Des informations complémentaires sont données dans la notice d'utilisation qui accompagne la macro.

La figure 6 donne le résultat de l'application de la macro aux poids des poissons. Comme signalé au paragraphe 2.2, on rejette l'hypothèse de normalité des données de départ au seuil $\alpha = 0,05$, la corrélation entre les observations et les scores normaux étant égale à 0,9334, alors que la valeur théorique est de 0,9870.

La transformation logarithmique, à laquelle on pense assez naturellement compte tenu de la dissymétrie gauche que présentent les observations, ne permet pas de normaliser les données : la corrélation est de 0,9582, soit une valeur inférieure à la valeur théorique.

La transformation racine carrée conduit également au rejet de l'hypothèse de normalité. Toutefois, la corrélation, égale à 0,9855, est relativement proche de la valeur théorique, égale à 0,9870. Rappelons que la valeur $\lambda = 0,5$ se situe dans l'intervalle de confiance de l'estimation de λ par la méthode du maximum de vraisemblance (paragraphe 2).

La transformation de BOX et COX, avec $\hat{\lambda} = 0,37$, réalise par contre la normalisation des données, de même que la transformation de JOHNSON. Pour cette dernière, la corrélation entre les données transformées et les scores normaux est encore plus élevée que pour la transformation de BOX et COX.

4.2. La macro SEQVARTRANSFO

Cette macro analyse successivement chacune des six séries définies au paragraphe 4.1, en arrêtant la recherche d'une transformation dès qu'une transformation pour laquelle l'hypothèse de normalité n'est pas rejetée a été rencontrée. L'objectif poursuivi est donc de trouver une transformation aussi simple que possible qui normalise les données.

Pour la transformation retenue, les données transformées peuvent être enregistrées et la même transformation peut être appliquée à des données supplémentaires. Les corrélations entre les données transformées et les scores normaux peuvent être enregistrées, à la fois pour la variable initiale et pour les données supplémentaires pour les différentes transformations qui ont été testées.

Des informations complémentaires sont reprises dans la notice d'utilisation qui accompagne le programme.

L'application de la macro aux données relatives aux amplificateurs montre que la première transformation qui normalise les données est la transformation logarithmique (figure 7). Pour cette transformation, la corrélation avec les scores normaux est égale au seuil de signification de 0,05.

```

%ALLVARTRANSFO c1

Valeurs théoriques des corrélations entre observations et
scores normaux :

Nobs =    99  cor10 = 0.9893  cor05 = 0.9870  cor01 = 0.9816
Alpha = 0.05  Ctheo = 0.9870

1. Données initiales: COR= 0.9334

2. Transf. Log.: COR= 0.9582

3. Transf. Box et Cox (Lambda= -1, 0.5, 0 ou .5):
   Lambda = 0.50  COR = 0.9855

4. Transf. Box et Cox (Lambda= -2 ou 2): Lambda =2.00  COR = 0.7596

5. Transf. Box et Cox, Lambda variant de -3.00 à 3.00 par pas de 0.10
   Lambda = 0.37 (Lim. conf. = 0.18 et 0.58)  COR = 0.9882

   --- Les données transformées XT = x**Lambda sont normales ---

6. Transf. de JOHNSON: Type = Su  COR = 0.9947
   ETA = 1.29317  GAMMA = -1.27776  LAMBDA = 43.82611
   EPSILON = 37.01955

   -- Les données transformées par la méthode de JOHNSON sont normales --

```

Figure 6 – Poids des poissons : commandes et résultats de la macro ALLVAR-TRANSFO.

Par rapport à la transformation de JOHNSON examinée au paragraphe 3, la valeur de la corrélation avec les scores normaux est plus faible : pour la transformation logarithmique on obtient 0,9891 (figure 7) alors que pour la transformation de JOHNSON on avait obtenu 0,9989 (figure 4).

5. CONCLUSIONS

Les macros proposées assurent une automatisation plus poussée de la recherche d'une transformation conduisant à des résultats pour lesquels l'hypothèse de non-normalité n'est pas rejetée que ne le permettent les commandes originales du logiciel Minitab. On notera cependant que des différences peuvent s'observer dans les résultats des macros proposées par rapport aux commandes de Minitab. Celles-ci sont dues aux algorithmes de calcul utilisés. Ainsi, pour la transformation de JOHNSON, Minitab recherche la transformation qui maximise la probabilité associée au test d'ANDERSON-DARLING, alors que la macro propo-

```

%SEQVARTRANSFO c10;
  XSUPPL c11;
  XSTRANS C12.
NAME c12 'XsTans'
PRINT c11-c12

Valeurs théoriques des corrélations entre observations et
scores normaux :

Nobs =   120  cor10 = 0.9910  cor05 = 0.9891  cor01 = 0.9846
Alpha = 0.05  Ctheo = 0.9891

1. Données initiales: COR= 0.9815

2. Transf. Log.: COR= 0.9891

-- Les données transformées XT = LOGE(X) sont normales --

Row      Xs    XsTans
1       7.75  2.04769
2      10.00  2.30259
3      12.25  2.50553

```

Figure 7 – Amplificateurs : commandes et résultats de la macro SEQVARTRANSFO.

sée maximise la corrélation entre les données transformées et les scores normaux. Cette petite nuance dans le choix du critère peut conduire à des ajustements assez différents, car on a vu que de nombreuses solutions conduisent à des résultats très proches du point de vue de la normalisation (figure 5).

La macro ALLVARTRANSFO analyse, de façon systématique plusieurs transformations de variables : les transformations classiques correspondant à des valeurs particulières du λ de la transformation de BOX et COX, la transformation de BOX et COX avec optimisation de λ et la transformation «optimale» de JOHNSON. A l'issue de cette analyse, l'utilisateur doit effectuer un choix. La question est dès lors la suivante : faut-il privilégier la transformation qui maximise la corrélation entre les données transformées et les scores normaux ou bien faut-il, comme le propose implicitement la macro SEQVARTRANSFO, privilégier des transformations plus simples ?

La réponse à la question n'est pas évidente et dépend du contexte. Si on souhaite maximiser la corrélation entre les données transformées et les scores normaux, la transformation de JOHNSON sera évidemment le plus souvent retenue, puisqu'elle est fonction d'un plus grand nombre de paramètres. Si on dispose d'un nombre d'observations suffisamment grand - de l'ordre de la centaine - et s'il est important d'avoir un ajustement de qualité, afin de calculer des quan-

tiles par exemple, la transformation de JOHNSON peut être indiquée. Ce sera le cas, par exemple, pour les problèmes d'étude de l'aptitude des procédés dans le contrôle statistique de la qualité.

Si on dispose de moins de données et si la normalité est simplement requise comme condition d'application de méthodes d'inférence, une transformation classique sera le plus souvent largement suffisante si elle conduit au non-rejet de l'hypothèse de normalité.

BIBLIOGRAPHIE

- BOX G. E. P., COX D. R. [1964]. An analysis of transformations (with discussion). *J. R. Stat. Soc.*, Ser B. 26(2), 211-252.
- CHOU Y. M., POLANSKI A. M., MASON R. L. [1998]. Transforming non-normal data to normality in statistical process control. *J. Qual. Technol.* 30 (2), 133-141.
- DAGNELIE P. [2006]. *Statistique théorique et appliquée. Tome 2 : Inférence statistique à une et à deux dimension*. Bruxelles, De Boeck et Larcier, 734 p.
- DRAPER N. R., SMITH H. [1998]. *Applied regression analysis*. New York, Wiley, 706 p.
- JOHNSON N. L. [1949]. Systems of frequency curves generated by methods of translation. *Biometrika* 36, 149-176.
- PALM R. [2002]. Macros Minitab pour la régression linéaire, 18 p.
<<http://www.fsagx.ac.be/si/>>.
- PALM R., IEMMA A. F. [2002]. Conditions d'application et transformations de variables en régression linéaire. *Notes Stat. Inform.* (Gembloux) 2002/1, 34 p.
- PEARN W. L., LIN P. C., CHEN K. S. [2004]. The C''_{pk} index for asymmetric tolerances : implications and inference. *Metrika* 60, 119-136.
- WEISBERG S. [2005]. *Applied linear regression*. New York, Wiley, 310 p.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant de l'Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie de l'Université de Liège/Gembloux Agro-Bio Tech et du Département Agriculture et milieu naturel, Unité Systèmes agraires, Territoire et Technologies de l'Information du Centre wallon de Recherches agronomiques.

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

*Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
Avenue de la Faculté d'Agronomie, 8
B-5030 GEMBLoux (Belgique)
E-mail : sima.gembloux@ulg.ac.be*

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

<http://www.fsagx.ac.be/si/>

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

- PALM R. [2007]. L'analyse des correspondances multiples : principes et application. *Notes Stat. Inform.* (Gembloux) 2007/2, 28 p.
- PALM R. [2008]. Détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée selon la norme ISO 5725-2. *Notes Stat. Inform.* (Gembloux) 2008/1, 22 p.
- CHARLES C., LECHARLIER L., RENAUD F. [2008]. Introduction à LATEX. *Notes Stat. Inform.* (Gembloux) 2008/2, 21 p.
- CHARLES C. [2008]. Introduction à OCTAVE. *Notes Stat. Inform.* (Gembloux) 2008/3, 19 p.
- PALM R., BROSTAUX Y. [2009]. Etude des séries chronologiques par les méthodes de décomposition. *Notes Stat. Inform.* (Gembloux) 2009/1, 17 p.
- CHARLES C. [2011]. Introduction aux ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/1, 22 p.
- CHARLES C. [2011]. Introduction aux applications des ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/2, 35 p.