REVUE DE STATISTIQUE APPLIQUÉE

G. CARLETTI

Détection automatique de valeurs anormales

Revue de statistique appliquée, tome 24, n° 3 (1976), p. 61-70

<http://www.numdam.org/item?id=RSA_1976__24_3_61_0>

DÉTECTION AUTOMATIQUE DE VALEURS ANORMALES [1]

G. CARLETTI
Faculté des Sciences Agronomiques de l'État, Gembloux (Belgique)

1 – INTRODUCTION

L'augmentation croissante du volume des données à traiter par l'ordinateur rend de plus en plus difficile le contrôle manuel de leur validité et l'automatisation de l'enregistrement de ces données n'a fait qu'accélérer ce phénomène. Aussi, nous a-t-il semblé opportun d'étudier des méthodes statistiques permettant de "filtrer" les données par un contrôle automatique. Dans cet article, on s'est limité à l'étude du problème à une dimension, c'est-à-dire en ne considérant qu'une seule variable à la fois.

Après avoir précisé deux méthodes de détection des valeurs anormales (paragraphe 2), nous aborderons la comparaison de différents tests de normalité (paragraphe 3). Nous utiliserons une transformation de variables de distributions quelconques en variables normales (paragraphe 4) et nous discuterons de l'organisation du programme de calcul (paragraphe 5) et de son application à un cas pratique (paragraphe 6). Nous donnerons ensuite quelques conclusions (paragraphe 7).

2 – METHODES DE DETECTION DES VALEURS ANORMALES

Nous avons choisi deux méthodes de détection des valeurs anormales qui avaient l'avantage d'utiliser toutes les observations contenues dans l'échantillon par l'intermédiaire des calculs de la moyenne et de l'écart-type et qui tenaient compte de l'effectif de l'échantillon.

2.1. – Méthode classique

La comparaison d'une valeur apparemment anormale x_i par rapport à l'ensemble des autres observations est identique à la comparaison de la

---

(*) Recherche subventionnée par l'Institut pour l'Encouragement de la Recherche Scientifique dans l'Industrie et l'Agriculture (I.R.S.I.A.).

Article remis en Décembre 1975, révisé en Mars 1976

Mots-clés : **Valeurs anormales, Tests de normalité, Transformation de variables, Programme de calculs.**

moyenne d'un échantillon d'une observation (l'observation anormale) avec la moyenne d'un échantillon formé des n – 1 autres observations (DAGNELIE, 1970).

On a :

$$T_{obs} = \frac{| x_i - \bar{x}' |}{\sqrt{\frac{SCE'}{n-2}\left(1 + \frac{1}{n-1}\right)}} = \frac{| x_i - \bar{x}' |}{\sqrt{\frac{n \quad SCE'}{(n-1)(n-2)}}} \qquad (1)$$

si on désigne par $\bar{x}'$ et SCE' la moyenne et la somme des carrés des écarts relatives aux n – 1 observations considérées comme normales.

On considère l'observation comme anormale lorsque :

$$T_{obs} \geqslant T_{1-\alpha'/2} = T_{1-\alpha/2n}$$

avec n – 2 degrés de liberté, le risque α' tenant compte du nombre d'observations. En effet, la probabilité de voir figurer une observation anormale, c'est-à-dire nettement inférieure ou supérieure à la moyenne, est d'autant plus élevée que l'effectif de l'échantillon est grand.

Pour que globalement, pour l'ensemble des n observations indépendantes constituant un échantillon aléatoire et simple, la probabilité de devoir écarter un résultat soit égale à α, il faut effectuer le test individuel, pour une observation, à un niveau α' approximativement égal à α/n (DAGNELIE, 1970).

2.2. – Méthode de GRUBBS.

On peut aussi calculer une seule fois la moyenne $\bar{x}$ et l'écart-type estimé de l'échantillon des n observations (y compris la valeur apparemment anormale x_i) et déterminer un T observé :

$$T_{obs} = \frac{| x_i - \bar{x} |}{s} \qquad (2)$$

avec

$$s = \sqrt{\frac{SCE}{(n-1)}}$$

et on considère que l'observation x_i est anormale lorsque :

$$T_{obs} \geqslant T_G$$

Les valeurs T_G de GRUBBS (1972) sont données dans une table pour des effectifs de 3 à 147 et pour des risques "globaux" de 1ère espèce de 0,1 %; 0,5 % ; 1 % ; 2,5 % ; 5 % ; 10 % (= $\alpha/2$ car les risques sont indiqués pour un test unilatéral). Ces valeurs peuvent être calculées en fonction des valeurs critiques de la distribution T de Student (GRUBBS, 1950 et 1969). En effet, on a la relation suivante :

$$T_G = \sqrt{\frac{n-1}{n} \frac{(n-1)\,T^2}{(n-2) + T^2}} \qquad (3)$$

où n est l'effectif de l'échantillon et T les valeurs $T_{1-\alpha/2n}$ (voir paragraphe

2.1. pour l'explication du $\alpha/2$ n) de Student avec un nombre de degrés de liberté égal à n – 2.

Dans le but d'isoler l'effet "valeur anormale" de toute cause de non normalité, la relation (2) sera parfois utilisée avec des observations et des paramètres transformés selon la technique utilisée par DRAPER et COX (voir paragraphe 4).

2.3. Calcul des valeurs T_G de GRUBBS à partir de la loi normale réduite

Le calcul des valeurs théoriques T_G de GRUBBS nécessite l'emploi de tables de STUDENT très complètes (SMIRNOV, 1961) pour des α' très faibles puisque égaux à α/n. En effet, pour un α global de 0,05,

$$T_{1-\alpha/2n} = T_{0,99750} \qquad \text{pour } n = 10$$

et

$$T_{1-\alpha/2n} = T_{0,9999750} \qquad \text{pour } n = 1000.$$

Cela nous a conduit à employer une méthode empirique de calcul :

$$T_G \simeq T_U = \frac{n-2}{n} \cdot U_{1-\alpha/2n} \tag{4}$$

cette relation a l'avantage d'utiliser les valeurs de U correspondant à la fonction de répartition de la distribution normale réduite et donne des résultats excellents à partir de $n > 10$. La relation (4) n'est cependant valable que pour un α global de 0,05, risque de première espèce le plus généralement utilisé. Le tableau 1 donne pour différents effectifs les valeurs T_G de GRUBBS (calculées par GRUBBS selon la relation (3), les valeurs T_U calculées par l'approximation normale (relation (4)) et les erreurs relatives commises sur T_G.

Tableau 1

Valeurs T_G de GRUBBS, valeurs T_U calculées par l'approximation normale et erreurs relatives commises sur T_G de GRUBBS exprimées en pour-cent ($\alpha = 0,05$; test bilatéral).

n	T_G	T_U	$100 \left\lvert \frac{T_U - T_G}{T_G} \right\rvert$
10	2,290	2,246	1,92
15	2,549	2,544	0,20
20	2,709	2,721	0,44
25	2,822	2,843	0,74
30	2,908	2,934	0,89
40	3,036	3,066	0,99
50	3,128	3,159	0,99
60	3,199	3,230	0,97
80	3,305	3,335	0,91
100	3,383	3,411	0,83
120	3,444	3,470	0,75
140	3,493	3,509	0,46

3 – COMPARAISON DE TESTS DE NORMALITE

Avec l'emploi de tests T ou T_G de GRUBBS, on pose une hypothèse de normalité de l'échantillon étudié. Pour vérifier la normalité, nous avons extrait de la littérature cinq tests dont le principe est de comparer un paramètre observé à la valeur théorique attendue pour une distribution normale et qui avaient été appréciés pour différentes formes de distributions (SHAPIRO et al., 1968 ; MORICE, 1972 ; SNEYERS, 1974). Pour le test de GEARY, le paramètre considéré est le rapport entre l'écart moyen absolu et l'écart-type. Pour le test de DAVID, c'est le rapport entre l'amplitude et l'écart-type. Pour le test de d'AGOSTINO, c'est une quantité dépendant du rang de chaque observation et de l'écart-type. Pour les deux tests de PEARSON, il s'agit du coefficient de symétrie g_1 et du coefficient d'aplatissement b_2 :

$$g_1 = \frac{m_3}{s^3} \tag{5}$$

avec s l'écart-type et m_3 le moment centré d'ordre 3,

$$b_2 = \frac{m_4}{s^4} \tag{6}$$

avec s l'écart-type et m_4 le moment centré d'ordre 4.

Les cinq tests ont été appliqués à 184 variables de natures différentes, comprenant des données dendrométriques (*) et des résidus de regression. Les échantillons étaient d'effectifs compris entre 50 et 500,et certains étaient des sous-échantillons destinés à observer l'influence de l'effectif sur des variables identiques. Le tableau 2 donne les pourcentages d'échantillons jugés "non normaux" par chacun des tests au niveau $\alpha = 0{,}05$. La dernière colonne reprend les résultats pour le regroupement du test g_1 avec le test b_2.

On constate que la puissance de tous les tests de normalité diminue logiquement en même temps que l'effectif. Le test du coefficient b_2 de PEARSON reste cependant le plus solide à cet égard. Les tests g_1 et b_2 de PEARSON, utilisés séparément, constituent un ensemble qui a permis de détecter à lui seul la presque totalité des distributions non normales. Ces résultats confirment ceux de la littérature (SNEYERS, 1974 et SHAPIRO et al, 1968) sauf en ce qui concerne le test de d'AGOSTINO. En effet, celui-ci ne semble pas supérieur aux tests de GEARY et DAVID, mais présente un autre inconvénient pratique : il demande la mise par ordre croissant des observations, ce qui nécessite une trop grande proportion de temps de calcul pour des effectifs de plusieurs centaines d'observations. Dans notre programme de détection des valeurs anormales (voir § 5), le test de la normalité de la variable étudiée est effectué par l'ensemble des tests g_1 et b_2 de PEARSON,ce qui réunit efficacité et rapidité de calcul.

(*) : données dendrométriques = caractéristiques quantitatives d'arbres ou de peuplements forestiers.

Tableau 2

Pourcentages d'échantillons jugés "non normaux" par chacun des tests de normalité pour différents types de variables et différents effectifs d'échantillons

Type de variables	Nombre d'échantillons	Effectif	GEARY	DAVID	d'AGOSTINO	g_1	b_2	ensemble $g_1 + b_2$
variables dendrométriques	35	300 à 400	31	57	29	86	34	91
Sous-échantillons des variables dendrométriques	35	50 à 70	20	29	11	26	29	40
résidus de régression	114	200 à 500	88	79	84	90	92	100
Total	184	50 à 500	64	65	66	77	69	87

4 – TRANSFORMATION EN VARIABLES NORMALES.

On a envisagé ensuite la transformation des variables non-normales en variables normales. Pour ce faire, on a utilisé la transformation puissance de BOX et COX (1964) qui fait intervenir un seul paramètre, ce qui demande moins de temps de calcul quand on travaille par approximations successives.

Si y est la variable initiale et z la variable transformée, on a :

$$z = \frac{y^{\lambda} - 1}{\lambda} \qquad \text{pour } \lambda \neq 0 \tag{7}$$

et

$$z = \log y \qquad \text{pour } \lambda = 0$$

La valeur de λ a été déterminée par approximations successives en utilisant la méthode de DRAPER et COX (1969), qui a pour objectif d'arriver à l'égalité suivante :

$$g_1 = \frac{1}{3} V\, g_2 \tag{8}$$

où g_1 est le coefficient de symétrie (voir relation 5), V le coefficient de variation de la variable transformée et g_2 le coefficient d'aplatissement que l'on calcule comme suit :

$g_2 = b_2 - 3$, b_2 étant défini par la relation (6).

Cette méthode, utilisée sur plusieurs échantillons de variables dendrométriques, conduit à des distributions nettement symétriques (g_1 tendant vers zéro), mais pas toujours normales (b_2 restant parfois nettement différent de la valeur théorique 3). Le tableau 3 reprend les moyennes et écarts-types estimés des paramètres g_1 et b_2 obtenus pour 35 échantillons de variables dendrométriques avant et après transformation.

Tableau 3

Moyennes et écarts-types estimés de g_1 et b_2 pour 35 échantillons de variables dendrométriques avant et après transformation.

Variables	g_1		b_2	
	Moyennes	Ecarts-types estimés	Moyennes	Ecarts-types estimés
initiales	0,460	0,420	3,023	0,691
transformées	– 0,030	0,045	2,452	0,426

Même si les coefficients b_2 s'écartent sensiblement de la valeur théorique 3 correspondant à une distribution normale, ce n'est cependant pas un inconvénient majeur puisque l'utilisation des distributions t dans les comparaisons de moyennes exige surtout une hypothèse de symétrie des distributions (RATCLIFFE, 1968). L'emploi de la méthode de DRAPER et COX se justifie donc pleinement.

5 – ORGANISATION DU PROGRAMME DE CALCUL.

La détection automatique des valeurs anormales a été effectuée sur ordinateur IBM 1130 par un programme de calcul fournissant un document final sur lequel on trouve la liste des valeurs supposées anormales. La figure 1 reprend l'ordinogramme de ce programme. Après avoir "testé" la normalité de l'échantillon étudié par l'ensemble des deux tests g_1 et b_2 de PEARSON (voir paragraphe 3), on effectue éventuellement la transformation puissance de BOX et COX (relation (7) et paragraphe 4) et on détecte les valeurs anormales par la statistique de GRUBBS (voir paragraphes 2.2. et 2.3.). Une fois toutes les observations contrôlées, on recommence le processus en considérant les observations restantes (c'est-à-dire celles non détectées comme anormales) comme un tout, et ainsi de suite jusqu'au moment où plus aucune valeur anormale n'est signalée. Un tel procédé se justifie car les valeurs anormales détectées peuvent en masquer d'autres par modification des paramètres de la variable et donc du contrôle lui-même.

Dans les cas où une transformation est nécessaire pour normaliser une variable, le programme effectuera la détection automatique à la fois sur la variable originale et sur la variable transformée. En effet, une certaine proportion de valeurs anormales peut entraîner une modification sensible des coefficients g_1 et b_2 (FERGUSSON, 1961 et GRUBBS, 1969) et par là le résultat des tests de normalité, la puissance de la transformation et donc de nouveau la détection elle-même. La variable originale est considérée alors comme normale.

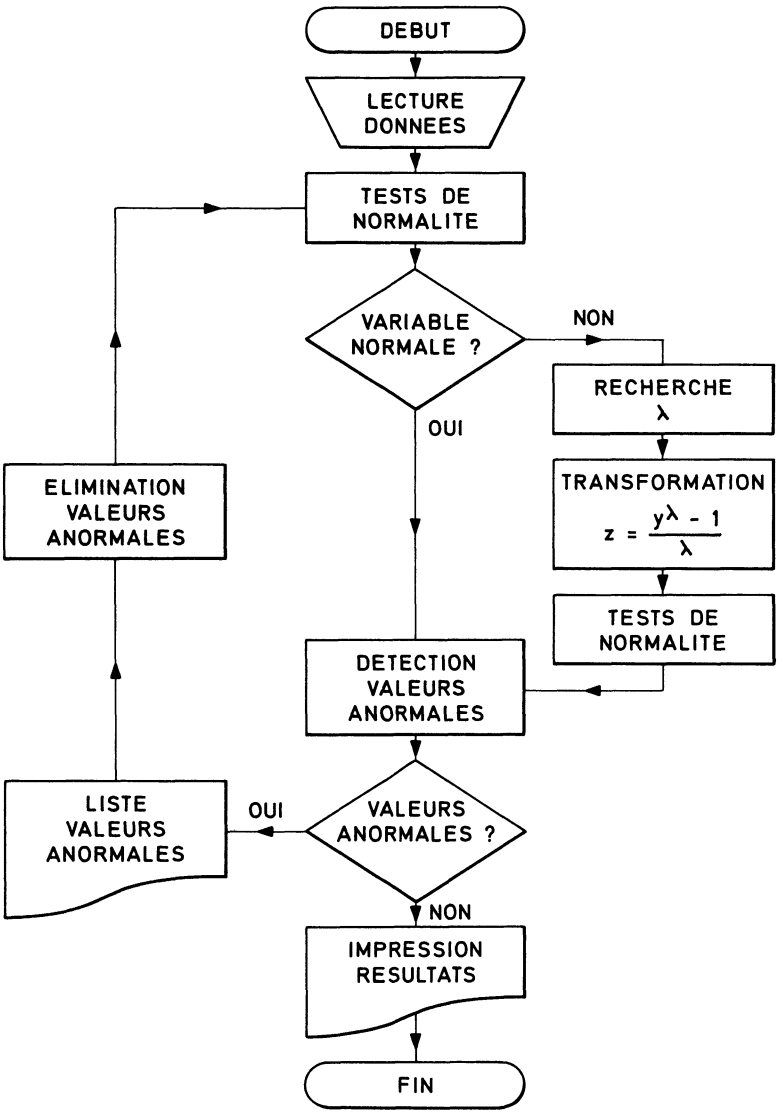

Figure 1 – Ordinogramme du programme de détection automatique des valeurs anormales.

6 – EXEMPLE D'APPLICATION.

L'échantillon étudié comme exemple est un ensemble de 434 mesures de circonférences (en cm) d'érables effectuées à 3,50 m de hauteur. Les tableaux 4 et 5 reprennent les paramètres de l'échantillon aux différentes étapes de la détection des valeurs anormales. La variable dont est tiré l'échantillon est considérée de distribution non normale lorsque les valeurs observées g_1 ou b_2 de l'échantillon sont supérieures aux fractiles théoriques $1 - \alpha = 0{,}95$ donnés par les tables des distributions de g_1 et de b_2 d'une distribution normale (indiqué par un astérisque). La recherche de λ pour la transformation $z = \dfrac{y^\lambda - 1}{\lambda}$ se fait par la méthode de DRAPER et COX (voir paragraphe 4).

Tableau 4

Paramètres de l'échantillon étudié pour les deux premières étapes de la détection des valeurs anormales.

Etapes	Effectif	Minimum	Maximum	Moyenne	Ecart-type	Coefficient de variation	g_1	b_2	λ
Echantillon initial	434	23	287	91,0	36,1	39,6	0,714*	4,561*	–
Echantillon transformé	434	6,60	23,94	13,43	2,745	20,4	– 0,002	2,851	0,4274

Pour la détection des valeurs anormales on compare les valeurs de T observé dans le T_G de GRUBBS. Les valeurs de T observé sont calculées sur l'échantillon transformé dans le but d'isoler l'effet "valeur anormale" de toute cause de non normalité. Soit le T_U de GRUBBS calculé par l'approximation normale (relation (4), paragraphe 2.3.) avec $\alpha = 0{,}05$ et $n = 434$:

$$T_G \cong T_U = \frac{n-2}{n} \cdot U_{1-\alpha/2n} = \frac{432}{434} U_{1-0{,}05/2.434} = 0{,}995\ U_{0{,}99994} = 3{,}82 \quad (9)$$

On calcule un T observé (relation (2), paragraphe 2.2.) pour chaque observation et on obtient une donnée anormale.

Soit cette observation $y = 287$, ce qui donne après transformation :

$$z = \frac{y^\lambda - 1}{\lambda} = \frac{287^{0{,}4274} - 1}{0{,}4274} = 23{,}94$$

et

$$T_{obs} = \frac{|\,23{,}94 - 13{,}43\,|}{2{,}745} = 3{,}83 \quad (10)$$

supérieur à la valeur théorique calculée en (9)

Après élimination de cette valeur anormale, le processus continue suivant l'ordinogramme de la figure 1. Le tableau 5 montre les modifications intervenues dans les paramètres de la variable étudiée. Les calculs effectués de manière similaire à (9) et à (10) ne détecteront plus de valeurs anormales dans les observations restantes.

Tableau 5

Paramètres de l'échantillon étudié après élimination de la donnée anormale.

Etapes	Effectif	Minimum	Maximum	Moyenne	Ecart-type	Coefficient de variation	g_1	b_2	λ
Echantillon initial	433	23	213	90,5	34,8	38,5	0,420*	2,939	–
Echantillon transformé	433	7,67	27,75	16,94	3,812	22,5	– 0,041	2,522*	0,505

7 – CONCLUSIONS.

On a précisé au cours des paragraphes 2.2. et 2.3. la méthode utilisée pour la détection automatique des valeurs anormales. Dans les paragraphes 5 et 6, on en a montré l'utilisation pratique. Il faut souligner que les variables "à contrôler" sont stockées sur support magnétique et ne sont pas modifiées automatiquement par ce processus. Les éliminations successives ont lieu en mémoire centrale dans des vecteurs de travail prévus pour cela et ont pour seul effet de signaler les valeurs qui paraissent anormales. En fin de compte, c'est au responsable qu'il appartiendra de vérifier et de juger si telle valeur signalée est bien anormale (impossibilité pratique, erreur de mesure...) et de l'éliminer alors réellement par une intervention au niveau du support magnétique. Cette méthode, qui a commencé à être mise en application dès 1974 au centre de calcul de la Faculté des Sciences Agronomiques de Gembloux, nous a convaincu de l'utilité de la détection automatique par sa rapidité et son efficacité pour de grands volumes de données où les contrôles manuels demeurent lents, imprécis ou quasi-impossibles.

8 – BIBLIOGRAPHIE.

BOX, G.E.P. et COX, D.R., An analysis of transformations. *J. Roy. Statist. Soc., Ser B.* (1964), 26, 211-252.

CARLETTI G., CLAUSTRIAUX J.J., DAGNELIE P., DEBOUCHE C., IN K., OGER R. et ROUSSEAUX G. Organisation d'une bibliothèque de programmes statistiques pour ordinateur. *Rev. Belg. Stat. Inf. Rech. Opér.* (1973), 12 (4), 2-16.

DAGNELIE P. Théorie et méthodes statistiques ; applications agronomiques (2 vol.). Presses Agron., Gembloux (1969-1970), 378 + 451 p.

DRAPER, N.R. et COX, D.R. On distributions and their transformation to normality. *J. Roy. Statist. Soc. série B.* (1969), 31, 472-476.

FERGUSSON T.S., Rules for rejection of outliers. *Rev. Inst. Int. de Stat.* (1961), 29 : 3, 29-43.

FISHER R.A. et YATES F. Statistical tables for biological, agricultural and medical research. Oliver and Boyd (1963), 146 p.

GRUBBS F.E., Testing outlying observations. *Ann. Math. Statist.* (1950), 21, 27-68.

GRUBBS F.E., Procedure for detecting outlying observations in samples. *Technometrics* (1969), 14, 1-21.

GRUBBS F.E. et BECK G. Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics* (1972), 14, 847-854.

MORICE E. Test de normalité d'une distribution observée. *Rev. Statist. Appl.* (1972), 20, 5-35.

PEARSON E.S. et HARTLEY H.O. Biometrika tables for statisticians (vol. I) University Press, Cambridge (1966), 264 p.

RATCLIFFE J.F. The effect on the t distribution of non-normality in the sampled population. *Appl. Statist.* (1968), 17, 42-48.

SHAPIRO S.S., WILK M.B. et CHEN H.J. A comparative study of various tests for normality. *J. Amer Statist. Ass.* (1968), 63, 1343-1372.

SMIRNOV N.V. Tables for the distribution and density functions of t-distributions. Pergamon, Oxford (1961), 129 p.

SNEYERS R. Sur les tests de normalité. *Rev. Statist. Appl.* (1974), 22, 29-36.