REVUE DE STATISTIQUE APPLIQUÉE

M. NIKULIN
A. ZERBET

Détection des observations aberrantes par des méthodes statistiques

Revue de statistique appliquée, tome 50, n° 3 (2002), p. 25-51
<http://www.numdam.org/item?id=RSA_2002__50_3_25_0>

Rev. Statistique Appliquée, 2002, **L** (3), 25-51

DÉTECTION DES OBSERVATIONS ABERRANTES PAR DES MÉTHODES STATISTIQUES

M. NIKULIN*, A. ZERBET*

* *Labotaroire de Statistique Mathématique et ses Applications, Université Victor Segalen Bordeaux 2, 146, rue Léo Saignat, B. P. 26, 33076 Bordeaux Cedex, France.*

RÉSUMÉ

Nous présentons ici un point de vue moderne sur le problème de détection des observations aberrantes, en particulier sur la règle de Chauvenet appliquée au modèle normal. A la base de cette règle, Bol'shev a proposé un test pour déceler simultanément plusieurs "outliers". D'un point de vue pratique, on a écrit un programme du test de Bol'shev, en langage Fortran, dans le but de détecter plusieurs "outliers" parmi les données expérimentales. Comme exemple, nous appliquons ce test aux données de Chauvenet (1863).

Mots-clés : *Détection, loi normale, loi de Thompson, inégalité de Bonferroni, observations aberrantes, "outlier", statistique exhaustive, règle de Chauvenet, test de Bolshev.*

ABSTRACT

We present here a modern point of view on outliers detection problem, in particular on Chauvenet's rule applied to the normal model. Based on this rule, Bol'shev proposed a test to detect simultaneously several outliers. From a practical point of view, we created a program of the Bol'shev's test in Fortran, in order to discover several outliers among the experimental data. As an example, we apply this test to Chauvenet's data (1863).

Keywords : *Detection, Bol'shev 's test, Bonferroni's inequality, Chauvenet's rule, exhaustive statistic, normal law, outliers, Thompson law.*

I. – Introduction

Le problème de la détection des observations aberrantes est universel et très ancien. Il se pose souvent à tous ceux qui ont à analyser des données expérimentales. Par exemple, supposons qu'après l'expérience, on ait vingt données que l'on suppose provenir d'une loi normale standard, que dix neuf d'entre elles, sont dans l'intervalle $[-4, 4]$ et que la vingtième ait la valeur 100000. La probabilité de rejeter cette dernière valeur est proche de 1, et on est presque certain qu'on n'aura pas d'erreur en rejetant cette valeur qui s'éloigne beaucoup des autres. Les premières règles de ce problème de détection sont proposées par Chauvenet en 1863. Puis pour le modèle normal, le problème est repris vers les années 1930 par plusieurs auteurs, dont Pearson et Chandra-Sekar (1936) qui ont proposé le test de détection dans le cas unilatéral (la présence d'une observation aberrante d'un seul côté, à droite ou à gauche). Mais ce

test peut être inefficace si l'échantillon contient plus d'une seule valeur dite "outlier". En 1950 Grubbs a présenté une règle de rejet d'au moins une valeur douteuse, dans les deux cas, unilatéral et bilatéral. Wilks en 1963 a étendu l'étude au cas général où il y a exactement k ($k \leqslant n-2$ (n le nombre des données expérimentales)) observations qui contiennent de grosses erreurs. Nous soulignons également ici l'importance du test de Bol'shev (1969) qui ne suppose pas que l'on connaisse le nombre exact des observations aberrantes, mais seulement qu'il ne dépasse pas un nombre maximum s, contrairement aux autres tests, comme ceux de Pearson et Chandra Sekar, Grubbs et Wilks qui supposent à l' avance que l'on connaisse ce nombre exact. En 1974, Bol'shev et Ubaidulaeva ont étudié le comportement asymptotique de la distribution du nombre N d'observations aberrantes par la règle de Chauvenet pour la famille des lois normales, lorsque l'hypothèse nulle est vraie (la loi de N se comporte pour $n \longrightarrow \infty$ comme une loi de Poisson). Ces résultats ont été généralisés par Ibragimov et Khalfina (1978) à une classe de distributions $\{F\left(\frac{x-\mu}{\sigma}\right)\}$ dépendant de deux paramètres représentant l'espérance et l'écart-type, à condition que les distributions de cette classe aient une variance finie. On note que la distribution de Cauchy ne vérifie pas cette condition; par conséquent, le test de Bol'shev n'est pas valable pour cette loi. On remarque aussi que les résultats de Bol'shev sont facilement applicables au cas d'une loi normale multimensionnelle (voir Voinov and Nikulin, (1996)), et en conséquence dans l'analyse des variances pour le modèle de Gauss-Markov (voir Greenwood and Nikulin, (1996)).

L'intérêt essentiel de cet article réside :

Premièrement dans le test de Bol'shev basé sur la règle de Chauvenet, qui permet de déceler simultanément plusieurs observations aberrantes.

Deuxièment dans la démonstration du théorème de Bol'shev-Thompson, faite dans le cas multidimensionel; ainsi que dans la détermination des quantiles de certains tests de détection.

Cependant quand une valeur est dite "outlier", cela ne signifie pas que cette valeur est nécessairement sans importance. Au contraire, parfois elle présente le cas extrême d'un phénomène particulièrement intéressant. C'est pour cette raison qu'un groupe de statisticiens suggère de garder les valeurs aberrantes comme une partie intégrante de leurs échantillons pour éviter de perdre l'information correspondante.

2. – Motivations et définition

2.1. Motivations

Dans les problèmes liés à l'analyse statistique qui concerne le test d'une hypothèse H sur la nature de l'expérience, il est essentiel de connaître la qualité des données expérimentales, d'où l'importance de la détection des observations aberrantes, puisque leur présence peut être l'une des causes principales de rejet de cette hypothèse.

2.2. Définition

On ne peut pas donner une définition générale des erreurs grossières, qui sont souvent appelées en statistique "observations aberrantes", puisqu'elles dépendent du test de détection et de l'hypothèse nulle. Ces erreurs sont souvent dues à la lecture incorrecte sur l'appareil de mesure, à de mauvais calculs etc ... , et elles induisent donc des données erronées.

En général, ces observations diffèrent sensiblement et s'éloignent beaucoup des autres valeurs de l'échantillon dont elles sont issues.

2.3. Modèle mathématique de la présence des observations aberrantes

Considèrons n variables aléatoires $Z_1, Z_2, ..., Z_n$ de même loi, et soit z un nombre réel donné.

Notre problème est de tester l'hypothèse H :

$$\mathbf{P}\{Z_i < z\} = F(z), \quad i = 1, ..., n,$$

contre l'hypothèse alternative H^+, d'après laquelle :

$$\mathbf{P}\{Z_i < z\} = (1 - \varepsilon)F(z) + \varepsilon G(z), \quad i = 1, ..., n, \quad 0 < \varepsilon < \frac{1}{2}, \quad (1)$$

où F et G sont des fonctions de répartition données (Fig. 1) telles que

$$G(z) < F(z) \quad , \quad z \in \mathbb{R}.$$

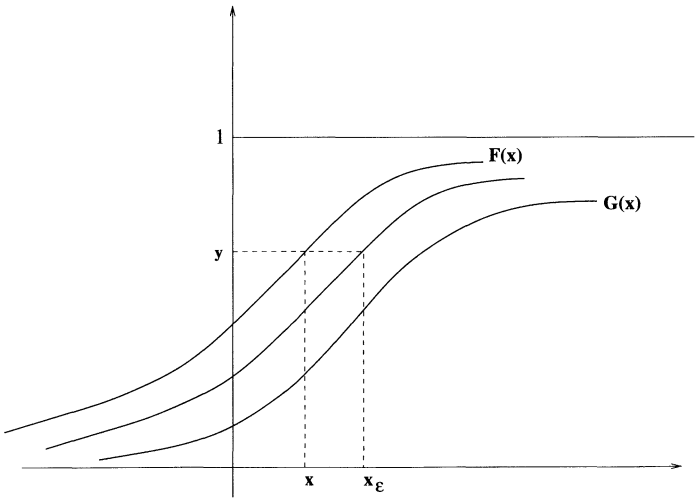

Figure 1

On prend $y \in]0,1[$ sur l'axe des ordonnées, et soit x et x_ε respectivement, la projection sur l'axe des abscisses de l'intersection de la droite qui passe par y et parallèle à l'axe des abscisses avec les courbes des fonctions de distribution F et $(1-\varepsilon)F + \varepsilon G$. Il est clair que d'après cette figure nous décelons des observations x_ε aberrantes à droite.

En développant cette procédure, nous en déduisons que pour détecter les valeurs douteuse à gauche, il faut tester H contre l'alternative H^-. Sous H^- la distribution des Z_i est donnée par la formule (1) avec :

$$G(z) > F(z) \quad , \quad z \in \mathbb{R}.$$

Pour le cas bilatéral, nous testons H contre l'alternative H^+_-. Sous cette alternative la distribution des Z_i est donnée par la formule (1) où la fonction $F - G$ change de signe une seule fois dans IR, cette fonction étant positive à $+\infty$ et négative à $-\infty$ (Fig. 2).

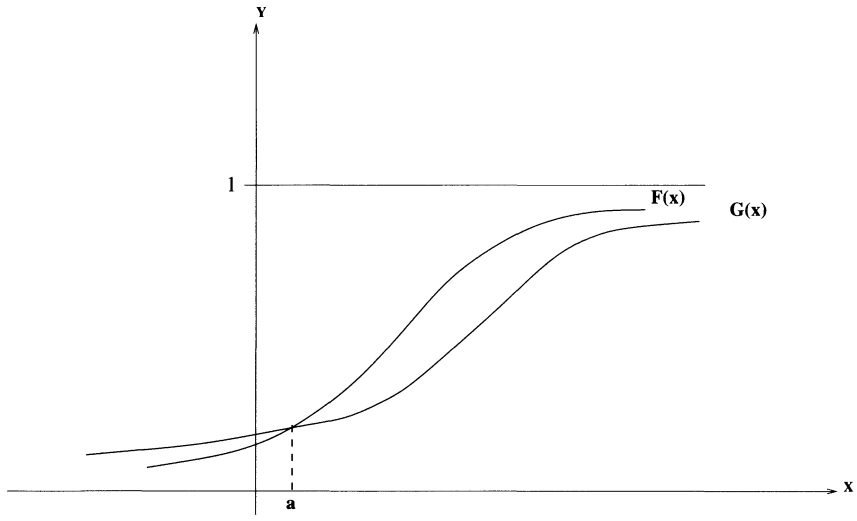

Figure 2

Soit a l'abscisse du point d'intersection des deux courbes correspondant à F et G. D'après ce qui précède, il est évident que dans $]a, +\infty[$ respectivement $]-\infty, a[$, on détecte des observations aberrantes respectivement à droite et à gauche.

2.4. Exposé du problème : cas classique de la famille normale

Nous avons un échantillon normal $\mathbf{X} = (X_1, X_2, ..., X_n)^T$, c'est-à-dire que $X_1, X_2, ..., X_n$ sont indépendamment distribuées de loi normale, d'espérances a_i, $i = 1, ..., n$ et variance σ^2. Nous voulons tester l'hypothèse $H_0 : X_i \sim N(a, \sigma^2)$,

$i = 1, 2, ..., n$ contre les trois hypothèses alternatives H_1^+, H_1^-, H_1 déterminées par les conditions suivantes :

$$a_1 = a_2 = ... = a_{m-1} = a_{m+1} = ... = a_n = a, \quad a_m = a + d,$$

avec respectivement pour :

$$H_1^+ : d > 0, \qquad H_1^- : d < 0, \qquad H_1 : d \neq 0.$$

Nous supposons que l'indice m de l'observation aberrante X_m, ainsi que la valeur d, sont inconnus.

Quand nous rejetons H_0 et qu'en réalité cette hypothèse est vraie nous commettons l'erreur de première espèce et nous appelons α le seuil de signification utilisé lors de l'élaboration du test statistique. Il est possible de le choisir aussi petit que nous voulons et de minimiser ainsi le risque de rejeter H_0 lorsqu'elle est vraie. Par contre, il existe aussi le risque β (erreur de deuxième espèce) d'accepter H_0 alors qu'en réalité elle est fausse. En pratique, il arrive fréquemment que l'espérance ou la variance ou les deux sont inconnues; nous devons alors les estimer par la méthode du maximum de vraisemblance. Sous l'hypothèse $H_0 : X_i \sim N(a, \sigma^2)$, $i = 1, 2, ..., n$, il y a quatre cas possibles : Les paramètres a et σ sont connus (cas <1.1>), l'espérance a est connue mais σ est inconnu (cas <1.0>), le cas contraire a est inconnu et σ est connu (cas <0.1>) et finalement le cas où les deux paramètres a et σ sont inconnus (cas <0.0>).

Dans ce modèle, la statistique

$$U = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right)^T \tag{2}$$

est exhaustive minimale et complète pour $\theta = (a, \sigma^2)^T$, et $\hat{\theta} = \left(\bar{X}_n, s_n^2\right)^T$ est l'estimateur du maximum de vraisemblance pour θ, où

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{et} \quad s_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2. \tag{3}$$

Nous pouvons construire les variables aléatoires Y_i, $i = 1, ..., n$, suivantes

$$Y_i = \begin{cases} \frac{1}{\sigma}(X_i - a), & \text{dans le cas <1.1>,} \\ \frac{1}{\sigma}\sqrt{\frac{n}{n-1}}(X_i - \bar{X}_n), & \text{dans le cas <0.1>,} \\ \frac{1}{s_n}(X_i - a), & \text{dans le cas <1.0>,} \\ \frac{1}{s_n}(X_i - \bar{X}_n), & \text{dans le cas <0.0>,} \end{cases} \tag{4}$$

avec

$$s_n^2 = \begin{cases} \frac{1}{n}\sum_{i=1}^n (X_i - a)^2, & \text{dans le cas } <1.0>, \\ \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2, & \text{dans le cas } <0.0>. \end{cases}$$

La distribution du vecteur aléatoire $\mathbf{Y}_k = (Y_1, ..., Y_k)$, $1 \leqslant k \leqslant n$ est indépendante des paramètres inconnus. La densité d'un tel vecteur aléatoire dans les quatre cas est donnée par Bol'shev et Ubaidullaeva (1974).

Dans le cas particulier où $k = 1$ la densité de Y_i, $i = 1, .., n$ coïncide avec celle de la loi normale standard dans les deux cas $<1.1>$ et $<0.1>$. Dans les autres cas $<1.0>$ et $<0.0>$ c'est la densité de Thompson respectivement à $f = n - 1$ et $f = n - 2$ degrés de liberté.

Présentons le théorème qui nous donne la densité de $\mathbf{Y}_k$ dans le cas $<0.0>$.

Dans le cas d'un échantillon normal de paramètres inconnus (cas $<0.0>$), on remarque que

$$\sum_{i=1}^n Y_i = 0 \quad , \quad \sum_{i=1}^n Y_i^2 = n. \tag{5}$$

La loi du vecteur $\mathbf{Y}_k = (Y_1, ..., Y_k)$ est dégénérée si $k = n - 1$ ou $k = n$, et sa densité est donnée par le théorème suivant pour $1 \leqslant k \leqslant n - 2$:

2.4.1. Théorème de Bol'shev-Thompson (Bol'shev(1969))

La loi de $\mathbf{Y}_k$ est absolument continue et sa densité $p(y_1, ..., y_k)$ est la suivante

$$p(y_1, ..., y_k) = \begin{cases} \frac{1}{(n\pi)^{\frac{k}{2}}} \sqrt{\frac{n}{n-k}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)} (1 - r^2)^{(n-k-3)/2}, & si \quad r < 1 \\ 0, & si \quad r \geqslant 1, \end{cases} \tag{6}$$

où

$$r^2 = \sum_{j=1}^k \frac{n - j + 1}{n(n-j)} \left(y_j + \frac{1}{n - j + 1} \sum_{i=1}^{j-1} y_i \right)^2 . \tag{7}$$

Cette loi s'appelle loi de Thompson à $n - k - 1$ degrés de liberté.

La démonstration du théorème ci-dessus est donnée en annexe.

En particulier pour $k = 1$ on a

$$p_{Y_1}(y) = \frac{1}{\sqrt{\pi(n-1)}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \left(1 - \frac{y^2}{n-1}\right)^{(n-4)/2}, \ |y| < \sqrt{n-1}. \tag{8}$$

qui est la loi de Thompson à $f = n - 2$ degrés de liberté.

On note

$$p_{Y_1}(y) = t_f(y) \quad , \quad T_f(y) = \int_{-\sqrt{f+1}}^{y} t_f(z)dz, \ |y| < \sqrt{f+1} \tag{9}$$

la densité et la fonction de répartition de la loi de Thompson à f degrés de liberté.

Il est facile de prouver les relations importantes ci-desous qui lient les fonctions de répartitions de Thompson et Student :

$$T_m(u_m) = S_m\left(\frac{u_m\sqrt{m}}{\sqrt{m+1-u_m^2}}\right) \quad et \quad S_m(v_m) = T_m\left(\frac{v_m\sqrt{m+1}}{\sqrt{m+v_m^2}}\right), \tag{10}$$

où S_m désigne la fonction de répartition de la loi de Student à m degrés de liberté; à l'aide de transformations non linéaires (10), on peut déduire la table des quantiles de la loi de Thompson à partir de celle de Student.

REMARQUE 1. – *On remarque que les variables aléatoires $Y_1, ..., Y_n$ ont la même distribution que celle de Thompson. Cette distribution ne dépend pas des paramètres a et σ^2. Pour cette raison nous pouvons prendre, par exemple, $a = 0$ et $\sigma^2 = 1$. Mais de (5) il s'ensuit que $Y_1, ..., Y_n$ sont dépendantes.*

2.5. Présentation de quelques tests de détection

2.5.1. Test de Dixon

Dixon (1950, 1951) a proposé la statistique de forme générale suivante :

$$D = \frac{X_{(q)} - X_{(p)}}{X_{(s)} - X_{(r)}}$$

où $1 \leqslant r \leqslant p < q \leqslant s \leqslant n$, $X_{(i)}$ étant la $i^{ème}$ statistique d'ordre.

Soit D_α la valeur critique du test au seuil α. Dans le cas unilatéral à droite (respectivement gauche), si la valeur de la statistique de test D est plus grande que la valeur D_α, l'observation $X_{(q)}$ (respectivement $X_{(p)}$) est alors aberrante.

Pour tester H_0 contre H_1^+, on utilise la statistique D pour $q = s = n, p = n-1$ et $r = 1$:

$$D = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}}.$$

Pour tester H_0 contre H_1^-, on utilise la statistique D pour $q = 2, p = r = 1$ et $s = n$:

$$D = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}.$$

2.5.2. Test de E. Pearson et Chandra-Sekar

E. Pearson et Chandra-Sekar (1936) ont proposé pour tester H_0 contre H_1^+ et H_1^-, les tests fondés sur les statistiques extrémales :

$$U_+ = \max_{1\leqslant i\leqslant n} Y_i = Y_{(n)} \quad \text{et} \quad U_- = \min_{1\leqslant i\leqslant n} Y_i = Y_{(1)}.$$

La région critique avec un risque de première espèce égal à α du test unilatéral à droite (respectivement gauche) est $]u_+^\alpha, +\infty[$ (respectivement) $]-\infty, u_-^\alpha[$), u_+^α (respectivement u_-^α) étant le quantile d'ordre $1-\alpha$ (respectivement α) de la loi de U_+ (respectivement U_-).

2.5.3. Test de Grubbs

Pour le cas bilatéral, Grubbs a proposé en 1950 le test de H_0 contre H_1, fondé sur la statistique

$$U = \max_{1\leqslant i\leqslant n} |Y_i| = \max\{Y_{(n)}, -Y_{(1)}\}.$$

Si u^α est le quantile d'ordre $1-\frac{\alpha}{2}$ de la statistique U, la région critique du test bilatéral, avec un risque de première espèce égale à α est : $]-\infty, -u^\alpha[\cup]u^\alpha, +\infty[$.

Il est clair (par symétrie) que les statistiques

$$U_+ \text{ et } U_-$$

suivent la même distribution. Pour cela nous proposons de n'étudier que les distributions de U_+ et U.

Fixons x positif, et soient les événements

$$A_i = \{Y_i \geqslant x\}, \quad B_i = \{|Y_i| \geqslant x\}, \quad i = 1, ..., n.$$

Il est évident, Y_i ayant une distribution symétrique, que

$$\mathbf{P}\{B_i\} = 2\mathbf{P}\{A_i\},$$

et pour tout $i = 1, 2, ..., n$, la probabilité de l'événement A_i est donnée par

$$\mathbf{P}\{A_i\} = \begin{cases} 1-\Phi(x), & \text{dans le cas <1.1>,} \\ 1-\Phi(x), & \text{dans le cas <0.1>,} \\ 1-T_{n-1}(x) = 1-S_{n-1}\left(x\sqrt{\frac{n-1}{n-x^2}}\right), & \text{dans le cas <1.0>,} \\ 1-T_{n-2}(x) = 1-S_{n-2}\left(x\sqrt{\frac{n-2}{n-1-x^2}}\right), & \text{dans le cas <0.0>,} \end{cases} \tag{11}$$

où S_m est la fonction de répartition de la loi de Student à m degrés de liberté, et Φ la fonction de répartition de la loi normale standard.

On sait que les événements A_i, A_j, et B_i, B_j sont deux à deux non positivement corrélés, i.e.

$$\mathbf{P}\{A_iA_j\} \leqslant \mathbf{P}\{A_i\}\mathbf{P}\{A_j\} \text{ et } \mathbf{P}\{B_iB_j\} \leqslant \mathbf{P}\{B_i\}\mathbf{P}\{B_j\},$$
$$i,j = 1,...,n \quad ; \quad (i \neq j).$$

Comme on le sait

$$\{U_+ \geqslant x\} = \cup_{i=1}^n A_i \quad \text{et} \quad \{U \geqslant x\} = \cup_{i=1}^n B_i,$$

Si $D_1, D_2, ..., D_n$ sont n événements, l'inégalité de Bonferroni du deuxième ordre pour ces événements est la suivante :

$$\sum_{i=1}^n \mathbf{P}\{D_i\} - \sum_{i<j} \mathbf{P}\{D_iD_j\} \leqslant \mathbf{P}\{\cup_{i=1}^n D_i\} \leqslant \sum_{i=1}^n \mathbf{P}\{D_i\}$$

L'inégalité de droite est celle de Bonferroni du premier ordre.

D'après l'inégalité ci-dessus, pour les événements respectivement $A_1, A_2, ..., A_n$ et $B_1, B_2, ..., B_n$ deux à deux non positivement corrélés, nous avons respectivement

$$-\frac{(n-1)p}{2} \leqslant \frac{\mathbf{P}\{U_+ \geqslant x\} - np}{np} \leqslant 0,$$

et

$$-(n-1)p \leqslant \frac{\mathbf{P}\{U \geqslant x\} - 2np}{2np} \leqslant 0,$$

où

$p = p(A_i)$ est défini par (11).

Pour les statistiques U_+ et U, nous pouvons donc prendre pour valeur critique x les solutions des équations

$$L : \begin{cases} \frac{Q}{n} = 1 - \Phi(x), & \text{dans le cas <1.1>,} \\ \frac{Q}{n} = 1 - \Phi(x), & \text{dans le cas <0.1>,} \\ \frac{Q}{n} = 1 - S_{n-1}\left(x\sqrt{\frac{n-1}{n-x^2}}\right) = 1 - T_{n-1}(x), & \text{dans le cas <1.0>,} \\ \frac{Q}{n} = 1 - S_{n-2}\left(x\sqrt{\frac{n-2}{n-1-x^2}}\right) = 1 - T_{n-2}(x), & \text{dans le cas <0.0>,} \end{cases}$$

dans ce cas les vrais niveaux de signification (risques de première espèce) des tests $\{U_+ \geqslant x\}$ et $\{U \geqslant x\}$ sont respectivement différents de Q et $2Q$ (au sens d'une erreur relative) de moins de $50Q\%$ et $100Q\%$.

Tabulations

On réalise, à l'aide de programmes Fortran, les tabulations des quantiles des statistiques U_+ et U, dans les différents cas <1.1>, <0.1>, <1.0> et <0.0> :

Quantiles des statistiques U_+ et U dans les deux cas <1.1> et <0.1>

	α				
n	0.01	0.02	0.05	0.1	0.2
5	2.87816	2.65207	2.32634	2.05375	1.75068
6	2.93520	2.71305	2.39397	2.12804	1.83391
7	2.98270	2.76374	2.44999	2.18935	1.90221
8	3.02334	2.80703	2.49770	2.24140	1.95996
9	3.05880	2.84476	2.53918	2.28655	2.00987
10	3.09023	2.87816	2.57582	2.32634	2.05375
11	3.11842	2.90810	2.60861	2.36189	2.09283
12	3.14397	2.93519	2.6382	2.39398	2.12804
13	3.16733	2.95994	2.66528	2.42319	2.16004
14	3.18881	2.98270	2.69010	2.44999	2.18935
15	3.20870	3.00375	2.71304	2.47473	2.21636
20	3.29052	3.09022	2.80703	2.57582	2.32634
25	3.35279	3.15591	2.87816	2.65206	2.40891
30	3.40293	3.20870	2.93520	2.71304	2.47473
35	3.44482	3.25276	2.98270	2.76374	2.52931

Quantiles des statistiques U_+ et U dans le cas <0.0>

	α				
n	0.01	0.02	0.05	0.1	0.2
5	1.95528	1.92892	1.86866	1.79068	1.66554
6	2.12981	2.08529	1.99602	1.89389	1.74645
7	2.26534	2.20519	2.09342	1.97444	1.81299
8	2.37416	2.30111	2.17192	2.04075	1.86984
9	2.46407	2.38040	2.23753	2.09720	1.91955
10	2.54006	2.44763	2.29377	2.14637	1.96372
11	2.60553	2.50576	2.34294	2.18992	2.00345
12	2.66278	2.55680	2.38655	2.22899	2.03952
13	2.71347	2.60219	2.42570	2.26439	2.07253
14	2.75883	2.64298	2.46118	2.29674	2.10293
15	2.79979	2.67996	2.49359	2.32651	2.13111
20	2.95873	2.82496	2.62299	2.44724	2.24703
25	3.07068	2.92863	2.71778	2.53741	2.33505
30	3.15594	3.00849	2.79206	2.60897	2.40565
35	3.22419	3.07301	2.85283	2.66806	2.46437

Quantiles des statistiques U_+ et U dans le cas <1.0>

	α				
n	0.01	0.02	0.05	0.1	0.2
5	2.11959	2.07073	1.97267	1.86028	1.69767
6	2.25344	2.18930	2.06999	1.94277	1.76975
7	2.36147	2.28487	2.14927	2.01135	1.83128
8	2.45112	2.36432	2.21594	2.07000	1.88489
9	2.52718	2.43198	2.27335	2.12118	1.93230
10	2.59287	2.49064	2.32365	2.16652	1.97475
11	2.65044	2.54227	2.36833	2.20718	2.01313
12	2.70150	2.58825	2.40848	2.24400	2.04812
13	2.74726	2.62962	2.44487	2.27761	2.08025
14	2.78862	2.66716	2.47812	2.30850	2.10992
15	2.82627	2.70146	2.50869	2.3370	2.13747
20	2.97494	2.83816	2.63240	2.45397	2.25133
25	3.08175	2.93768	2.72431	2.54218	2.33822
30	3.16405	3.01515	2.79691	2.61257	2.40811
35	3.23042	3.07816	2.85661	2.67090	2.46635

A l'aide des résultats expérimentaux, nous calculerons la statistique de test et nous déterminerons s'il se situe dans la région critique; si oui, on rejette l'hypothèse H_0; sinon nous acceptons H_0.

A partir d'une table de la loi de la statistique de test, nous pouvons trouver la valeur critique X_α correspondant au quantile α. La région critique du test respectivement unilatéral à droite et bilatéral est respectivement $K_+ =]X_\alpha, +\infty[$ et $K_- \cup K_+$ avec $K_- =]-\infty, -X_\alpha[$, avec les risques de première espèce égaux à α et 2α respectivement.

2.5.4. Exemple (Barnett et Lewis (1994), page 15)

Supposons qu'on ait $n = 10$ résultats expérimentaux de mesure de notre variable d'étude :

$$X_1 = 1.74, \quad X_2 = 1.46, \quad X_3 = -1.28, \quad X_4 = -0.02 \quad X_5 = -0.40,$$
$$X_6 = 0.02, \quad X_7 = 3.89, \quad X_8 = 1.35, \quad X_9 = -0.10, \quad X_{10} = 1.71,$$

l'hypothèse nulle H_0, que nous désirons tester, signifie que les résultats ci-dessus suivent la loi normale $N(a, 1)$ d'espérance a inconnue et variance égale à 1.

Pour le niveau de signification $\alpha = 0.05$, la valeur critique Y_c, en utilisant la table des quantiles dans le cas <0.1> pour $n = 10$, est

$$Y_c = 2.57582.$$

Comme critère de décision, on ne rejette l'hypothèse nulle que si $Y_i = \frac{X_i - \bar{X}}{1} > Y_c$.

On a,

$$\bar{X} = 0.837$$

et

$$Y_1 = 0.903, \quad Y_2 = 0.623, \quad Y_3 = -2.117 \quad Y_4 = -0.857, \quad Y_5 = -1.237,$$
$$Y_6 = -0.817, \quad Y_7 = 3.053, \quad Y_8 = 0.513, \quad Y_9 = -0.937, \quad Y_{10} = 0.873.$$

Pour tout $i \neq 7$, on trouve $Y_i < Y_c$, mais pour $i = 7$, on a $Y_7 = 3.053 > Y_c$. Donc la seule valeur qui contienne une grosse erreur est $X_7 = 3.89$, ce qui confirme le résultat obtenu dans (Barnett et Lewis (1994)).

REMARQUE 2. – *La répétition itérative de ces tests pour détecter successivement plusieurs valeurs douteuses n'est pas conseillée puisque dès qu'on rejette la première observation aberrante, les autres observations ne sont pas indépendantes; ainsi la puissance du test devient faible. On constate donc que ces tests sont utilisables seulement pour déceler une seule observation aberrante.*

2.6. Test de Wilks

Considérons le problème qui consiste à tester l'hypothèse H_0 contre l'hypothèse alternative H_k selon laquelle il y a exactement k ($k \leqslant n-2$) observations $Z_{i_1}, ..., Z_{i_k}$ pour lesquelles les moyennes sont différentes de a (on suppose que les rangs $i_1, ..., i_k$ de ces observations sont inconnus). Pour tester H_0 contre H_k, il est naturel d'utiliser le théorème de Bol'shev-Thompson pour construire un test fondé sur la statistique suivante :

$$\rho^2 = \max_{i_1,...,i_k} r^2(Y_{i_1}, ..., Y_{i_k}),$$

$r^2 = r^2(y_1, ..., y_k)$ est la forme quadratique donnée par :

$$r^2 = \sum_{j=1}^{k} \frac{n-j+1}{n(n-j)} \left[y_j + \frac{1}{n-j+1} \sum_{i=1}^{j-1} y_i \right]^2.$$

Il découle de la démonstration du théorème de Bol'shev-Thompson (cf. annexe) et de l'inégalité de Bonferroni que :

$$\mathbf{P}\{\rho^2 \geqslant R^2\} \leqslant \binom{n}{k} \mathbf{P}\{r^2(Y_1, ..., Y_k) \geqslant R^2\} = \binom{n}{k} I_{1-R^2}\left(\frac{n-k-1}{2}, \frac{k}{2}\right).$$

$I_y(r, s)$ étant la valeur en y de la fonction de répartition d'une loi Beta à r et s degrés de liberté. Ce test est équivalent au test proposé par Wilks (1963). Pour avoir la borne inférieure de la probabilité de l'événement $\{\rho^2 \geqslant R^2\}$, on se sert de l'inégalité de Bonferroni. C'est pourquoi on recommande de calculer

$$\binom{n}{k} I_{1-R^2}\left(\frac{n-k-1}{2}, \frac{k}{2}\right).$$

où R^2 est la valeur observée de ρ^2. Si sa valeur n'est pas supérieure au niveau de signification donné Q, alors H_0 doit être rejetée. Le niveau de signification exact de ce test n'est pas supérieur à celui qui est donné.

Notons que,

$$\binom{n}{k} I_{1-R^2}\left(\frac{n-k-1}{2}, \frac{k}{2}\right),$$

est le nombre moyen des combinaisons de k observations parmi n pour lesquelles $\rho^2 \geqslant R^2$. D'où il s'ensuit que Wilks dans son article a proposé de refuser le choix de la valeur critique R^2 d'après le niveau de signification donné et de s'orienter sur le nombre moyen donné.

2.7. La règle de Chauvenet

La règle de Chauvenet est un test ancien (1863) basée sur une propriété simple de l'espérance mathématique. Ce test détecte la présence d'au moins une valeur aberrante dans un ensemble de résultats de mesure et l'élimine.

Supposons que nous ayons comme dans la partie ci-dessus n variables aléatoires $Z_1, ..., Z_n$ de même loi et z un nombre réel donné, on détermine les variables aléatoires N_i et M_i, $(i = 1, ..., n)$ de la façon suivante :

$$N_i = \begin{cases} 1, & \text{si} \quad Z_i \geqslant z, \\ 0, & \text{sinon;} \end{cases} \qquad M_i = \begin{cases} 1, & \text{si} \quad |Z_i| \geqslant z, \\ 0, & \text{sinon.} \end{cases}$$

Alors les statistiques

$$N = \sum_{i=1}^{n} N_i \quad , \quad M = \sum_{i=1}^{n} M_i,$$

nous donnent le nombre d'observations parmi $Z_1, ..., Z_n$ qui sont supérieures au seuil z choisi, respectivement dans le cas unilatéral et bilatéral. Ils suivent respectivement une loi binomiale de moyenne

$$\mathbf{E}\{N\} = n\mathbf{P}\{Z_1 > z\},$$
$$\mathbf{E}\{M\} = n\mathbf{P}\{|Z_1| > z\}.$$

Si on suppose à l'avance que $\mathbf{E}\{N\} = \alpha$ est un nombre positif, alors la valeur critique $z(\alpha)$ est la solution de l'équation

$$\mathbf{P}\{Z_1 > z\} = \frac{\alpha}{n}.$$

Il est évident, si les variables $Z_1, ..., Z_n$ sont indépendantes et pour n assez grand et α suffisament petit, que

$$\begin{aligned}\mathbf{P}\{\max_{1<i<n} Z_i > z(\alpha)\} &= 1 - \{1 - \mathbf{P}\{Z_i > z(\alpha)\}\}^n \\ &= 1 - (1 - \frac{\alpha}{n})^n = 1 - e^{-\alpha} + o(1) \simeq \alpha,\end{aligned}$$

est le niveau de signification du test et il est approximativement α pour n grand et α petit.

Si de plus les observations $Z_1, ..., Z_n$ sont négativement corrélées, i.e.

$$\mathbf{P}\{Z_i \geqslant z, Z_j \geqslant z\} \leqslant \mathbf{P}^2\{Z_1 \geqslant z\},$$

en utilisant l'inégalité de Bonferroni :

$$-\frac{\mathbf{E}\{N\}}{2}\left(1 - \frac{1}{n}\right) \leqslant \frac{\mathbf{P}\{\max_{1\leqslant i\leqslant n} Z_i > z\} - \mathbf{E}\{N\}}{\mathbf{E}\{N\}} \leqslant 0$$

on estime l'erreur relative entre le vrai niveau de signification et $\mathbf{E}\{N\}$ à moins de $50\,\mathbf{E}\{N\}$ %. Chauvenet a suggéré de choisir z tel que $\mathbf{E}\{N\} = 1/2n$, autrement dit, supposons qu'on ait exactement une observation aberrante parmis $2n$ observations. Ces résultats sont vrais si les variables aléatoires Z_i, $i = 1, ..., n$ sont indépendantes. Cependant, lorsque elles ne sont plus indépendantes, on obtient des résultats analogues en particulier lorsque les $(Z_1, ..., Z_n)$ représentent une statistique exhaustive.

En réalité, toutes les mesures pour lesquelles Y_i a pris une valeur supérieure à la valeur critique $y(\alpha)$ doivent être considérées comme aberrante, mais en fait le test de Chauvenet nous permet seulement d'éliminer la plus grande mesure.

2.8. Test de Bol'shev

Soit à tester l'hypothèse H_0 d'après laquelle les variables aléatoires indépendantes $X_1, ..., X_n$ suivent la loi normale de paramètres a et σ^2 contre les hypothèses alternatives H_1^+ et H_1. Puisque les statistiques Y_i, $i = 1, ..., n$ ne sont plus indépendantes, il est raisonnable de chercher une autre statistique de test.

Etant donné un nombre positif t, soit $y(t)$ la solution de l'équation

$$T_{n-2}[y(t)] = 1 - \frac{t}{n}, \tag{12}$$

où T_{n-2} est la fonction de répartition de Thompson à $n-2$ degrés de liberté.

$y(t)$ est donc le quantile d'ordre $1 - \frac{t}{n}$ du la loi de Thompson à $n-2$ degrés de liberté.

Définissant pour tout Y_i, $N_i(t)$ et $M_i(t)$ par :

$$N_i(t) = \begin{cases} 1 & \text{si} \quad Y_i \geqslant y(t), \\ 0 & \text{sinon}, \end{cases} \qquad M_i(t) = \begin{cases} 1 & \text{si} \quad |Y_i| \geqslant y(t), \\ 0 & \text{sinon}. \end{cases} \tag{13}$$

Posons

$$N(t) = \sum_{i=1}^{n} N_i(t) \quad \text{et} \quad M(t) = \sum_{i=1}^{n} M_i(t) \tag{14}$$

le nombre d'observations qui sont supérieures au seuil $y(t)$ respectivement dans le cas unilatéral et bilatéral.

Pour tout entier $k \geqslant 1$, soit les nombres arbitraires $t_1, t_2, ..., t_k$ fixés tels que

$$0 \leqslant t_1 \leqslant ... \leqslant t_k \leqslant T.$$

Nous pouvons construire les vecteurs aléatoires $(N^1, ..., N^k)$ et $(M^1, ..., M^k)$,

$$N^1 = N(t_1), \quad M^1 = M(t_1), \quad N^i = N(t_i) - N(t_{i-1}),$$
$$M^i = M(t_i) - M(t_{i-1}), \quad i = 2, ..., k$$

THÉORÈME 1 (Bol'shev et Ubaidulaeva (1974)). – *Si* $n \longrightarrow \infty$, *alors pour toutes les constantes* $t_0, ..., t_k$ *telles que* $0 = t_0 < ... < t_k \leqslant T$ *et pour toute suite d' entiers positifs* $u_1, ..., u_k$

$$\mathbf{P}\{\bigcap_{i=1}^{k}[N^i = u_i]\} = \prod_{i=1}^{k} \frac{(t_i - t_{i-1})^{u_i}}{u_i!} e^{-(t_i - t_{i-1})} + o(1).$$

$$\mathbf{P}\{\bigcap_{i=1}^{k}[M^i = u_i]\} = \prod_{i=1}^{k} \frac{[2(t_i - t_{i-1})]^{u_i}}{u_i!} e^{-2(t_i - t_{i-1})} + o(1),$$

où l'entier $k \geqslant 1$. *Autrement dit les processus* $N(t)$ *et* $M(t)$ *pour* $0 \leqslant t \leqslant T$, *si* $n \longrightarrow \infty$ *convergent en probabilité vers les processus de Poisson* $L(t)$, *respectivement d'intensité* 1 *et* 2.

Notant que les variables aléatoires N^i et M^i, $i = 1, 2, ..., k$, sont indépendantes de la constante t_0.

COROLLAIRE 1. – *Soit* θ *fixé,* $\theta > 0$, *notons* $t(\theta)$ *le point du saut du processus* $N(t)$ *(ou* $M(t)$ *), tel que* $N(t) = \theta$, $N(t-0) = \theta - 1$ *(ou* $M(t) = \theta$, $M(t-0) = \theta - 1$ *). Dans ce cas la fonction* $H_\theta(t) = N(t)/\theta$, $0 < t \leqslant t(\theta)$ *(ou* $H_\theta(t) = M(t)/\theta$, $0 < t \leqslant t(\theta)$ *) converge en loi, si* $n \longrightarrow \infty$ *et* $t(\theta)$ *est fixé, vers la fonction de distribution empirique construite relativement à* θ *variables aléatoires mutuellement indépendantes, uniformement distribuées sur* $]0, t(\theta)]$.

Maintenant, on veut déterminer la valeur critique, correspondant au niveau de signification fixé α, à partir duquel on rejette les observations aberrantes.

Soit $L(t)$ un processus de Poisson d'intensité λ, $t > 0$, $\lambda > 0$, et x un nombre positif donné.

Notons

$$t_s = \frac{s}{[\lambda(1+x)]}, \quad s = 1, 2, \cdots,$$

et considérons les variables aléatoires

$$Z_s = Z_s(x) = \sup_{0<t<t_s} \frac{L(t) - \lambda t}{\lambda t}, \quad s = 1, 2, \cdots.$$

Puisque

$$\{Z_s \leqslant x\} = \bigcup_{i=0}^{s-1} \{Z_s \leqslant x, L(t_s) = i\},$$

alors

$$\mathbf{P}_s(x) = \mathbf{P}\{Z_s \leqslant x\} = \sum_{i=0}^{s-1} p_{si},$$

où $p_{si} = \mathbf{P}\{Z_s \leqslant x,\ L(t_s) = i\}$, ce qui est équivalent à

$$p_{si} = \sum_{j=0}^{i} \mathbf{P}\{Z_s \leqslant x,\ L(t_s) = i\ L(t_{s-1}) = j\}.$$

En utilisant les propriétés du processus de Poisson, il est facile de démontrer que les probabilités p_{si} sont liées par le système d'équations

$$p_{si} = \sum_{j=0}^{i} p_{s-1,j} \frac{(\lambda t_1)^{i-j}}{(i-j)!} \exp(-\lambda t_1), \quad i = 0, 1, \cdots, s-1; s = 2, 3, ..., \qquad (15)$$

avec $p_{ss} = 0$ et $p_{10} = \exp(-\lambda t_1)$.

La solution de ce système est exprimée par les expressions

$$p_{si} = (1 - \frac{i}{s}) \frac{(s\lambda t_1)^i}{i!} \exp(-s\lambda t_1), \quad s = 1, 2, ...,; \quad i = 0, 1, ..., s. \qquad (16)$$

De la formule (16) on déduit que

$$\mathbf{P}_s(x) = \sum_{i=0}^{s-1} p_{si} = \sum_{i=0}^{s-1} \frac{(s\lambda t_1)^i}{i!} \exp(-s\lambda t_1) - \lambda t_1 \sum_{i=0}^{s-2} \frac{(s\lambda t_1)^i}{i!} \exp(-s\lambda t_1).$$

En tenant compte de cette égalité et de la relation classique entre les fonctions de répartition de la loi de Poisson et de la loi Gamma qui s'écrit :

$$\sum_{i=0}^{s-1} \frac{(s\lambda t_1)^i}{i!} \exp(-s\lambda t_1) = 1 - J_{s\lambda t_1}(s),$$

on conclut que

$$\mathbf{P}_s(x) = 1 - J_{s\lambda t_1}(s) - \lambda t_1[1 - J_{s\lambda t_1}(s-1)], \tag{17}$$

où $J_y(m)$ est la fonction de répartition de la loi Gamma de paramètre m

$$J_y(m) = \frac{1}{\Gamma(m)} \int_0^y t^{m-1} \exp(-t)\, dt, \quad y > 0, \quad t > 0. \tag{18}$$

Pour tout entier positif s, et α un niveau de signification donné, nous définissons $b = b(s, \alpha)$ comme solution de l'equation $\mathbf{P}_s(x) = 1 - \alpha$, soit :

$$J_b(s) + \frac{b}{s}[1 - J_b(s-1)] = \alpha, \quad \alpha \in]0, \frac{1}{2}[.$$

Posons $x = \frac{s}{b} - 1$. Alors avec la probabilité $1 - \alpha$, $L(t) \leqslant \lambda t(1 + x) = \lambda st/b$ pour tout t dans le demi intervalle $0 < t \leqslant t_s = \frac{b}{\lambda}$; la propriété indiquée ci-dessus sera utilisée dans la construction du test pour la détection des observations aberrantes en nombre ne dépassant pas s.

En effet, on observe n variables aléatoires $X_1, X_2, ..., X_n$; celles qui correspondent à de grosses erreurs parmi ces observations sont celles pour lesquelles la valeur correspondante de $L(t)$ est plus grande que $\lambda st/b$.

REMARQUE 3. – *L'avantage du test de Bol'shev est qu'on ne suppose pas que l'on connaisse d'avance le nombre d'observations aberrantes contrairement aux tests de "Pearson et Chandra-Sekar", "Grubbs" et "Wilks", mais seulement qu'il ne dépasse pas un nombre maximum s. Ainsi, il détecte simultanément plusieurs "outliers".*

A cause de la complexité des calculs, Bol'shev a proposé une autre statistique de test

$$\tau = \min\{\frac{\tau_1}{1}, ..., \frac{\tau_s}{s}\},$$

où $\tau_1, ..., \tau_s$ sont les s sauts du processus de Poisson $L(t)$ d'intensité λ.

Il est facile de démontrer, compte tenu de (17), que

$$\mathbf{P}\{\tau \leqslant c\} = J_{\lambda cs}(s) + \lambda c[1 - J_{\lambda cs}(s-1)],$$

alors la valeur critique c est la solution de l'équation

$$J_{\lambda cs}(s+1) + \lambda c[1 - J_{\lambda cs}(s)] = \alpha$$

tel que α est le niveau de signification, $0 < \alpha < 0.5$.

On réalise, à l'aide d'un programme Fortran, la tabulation de $(\lambda c - \alpha)10^5$ pour différentes valeurs :

$$\alpha = (0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2) \quad \text{et} \quad s = 1(1)7,$$

on obtient les résultats suivants :

	α						
s	0.002	0.005	0.01	0.02	0.05	0.1	0.2
1	0.199771	1.25102	5.020919	20.2192	128.9931	534.598	2307.476
2	0.0005	0.0083	0.06618	0.52449	7.9859	61.6174	468.9562
3	0.000007	0.000068	0.001106	0.01737	0.64322	9.44789	129.51789
4	0.000007	0.000007	0.000022	0.000648	0.05838	1.63945	40.8382
5	0.000007	0.000007	0.0000076	0.000022	0.0056838	0.30471	13.82914
6	0.000007	0.000007	0.000007	0.000007	0.000572	0.05925	4.9006
7	0	0.000007	0.000007	0.000007	0.000053	0.01189	1.792617

REMARQUE 4. – *D'après ces résultats, on constate que pour s assez grand $(\lambda c - \alpha)10^5$ tend vers 0 autrement dit $c \longrightarrow \alpha/\lambda$, c'est pourquoi Bol'shev a pris la valeur critique égale à α/λ.*

2.8.1. Les étapes d'application du test de Bol'shev

Etant donné un échantillon $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ normal, tout d'abord, on construit les statistiques $V_1, ..., V_n$ uniforme sur l'intervalle $[0, n]$, où

$$V_i = \begin{cases} n[1 - T_{n-2}(Y_i)], & \text{dans le cas unilatéral,} \\ n[1 - T_{n-2}(|Y_i|)], & \text{dans le cas bilatéral,} \end{cases} \qquad \text{pour } i = 1, ..., n,$$

après, en utilisant les statistiques $V_1, ..., V_n$, on construit le vecteur des statistiques d'ordre $V^{(.)} = (V_{(1)}, ..., V_{(n)})^T$, $V_{(1)} \leqslant ... \leqslant V_{(n)}$.

Ensuite, supposant α fixé ($0 \leqslant \alpha \leqslant 0.5$), pour tout $X_i, i = 1, ..., n$, on calcule $\dfrac{V_{j(i)}}{j(i)}$ où $j(i)$ est l'indice de $V_{(.)}$ correspondant à X_i.

Si $\dfrac{V_{j(i)}}{j(i)} \leqslant \dfrac{\alpha}{\lambda}$, on rejette alors l'observation X_i ($\lambda = 1$ dans le cas unilatéral; $\lambda = 2$ dans le cas bilatéral).

REMARQUE 5. – *On remarque que les étapes de ce test ne dépendent pas de s. Ainsi pour le cas particulier de $s = 1$ le test de Bol'shev coïncide avec les tests de "Pearson et Chandra-Sekar", "Grubbs" , "Smirnov" et "Wilks".*

2.8.2. Exemple (Barnett et Lewis (1994), page 38)

On étudie l'exemple de l'ensemble des données analysées par Peirce en 1852 et Chauvenet en 1863. Il est constitué de 15 observations, qui présentent le demi-diamètre vertical de la planète Venus, et qui ont été établies par Lt. Herndon en 1846,

-0.30	$+0.48$	$+0.63$	-0.22	$+0.18$
-0.44	-0.24	-0.13	-0.05	$+0.39$
$+1.01$	$+0.06$	-1.40	$+0.20$	$+0.10$

Nous examinons ces données pour déterminer lesquelles sont aberrantes. Nous supposons qu'elles proviennent d'un échantillon normal de paramètres inconnus.

Dans un premier temps, on utilise la technique graphique quantile-quantile (q-q), pour l'ajustement graphique des observations à la loi normale de paramètres inconnus. Un graphique q-q est un tracé des quantiles de l'ensemble des valeurs observées en fonction des quantiles de l'ensemble des valeurs théoriques. Les avantages de ce graphique sont les suivants :

- les dimensions d'échantillon ne sont pas obligatoirement égales.
- plusieurs distributions peuvent être simultanément testées. Et la présence des "outliers" peut aussi être détectée par ce diagramme q-q.

Notons que nous avons établi les diagrammes ci-dessous à l'aide du logiciel SPSS.

Le diagramme quantile-quantile gaussien des variables aléatoires X_i, $i = 1, .., 15$ (Fig. 3), confirme que l'observation $X_{13} = -1.40$ est significativement loin de la droite de Henry. Pour cette raison, on a pensé à éliminer cette observation.

Nous signalons que l'axe des ordonnées du diagramme q-q des résidus gaussien correspond aux valeurs gaussiennes théoriques et l'axe des abscisses aux valeurs observées. Le second diagramme (Fig.4) présente les résidus gaussiens (les valeurs observées - les valeurs gaussiennes théoriques) en fonction des valeurs observées.

On présente le diagramme q-q des autres observations en éliminant X_{13} sur la figure 5.

On voit que $X_{11} = 1.01$ est aussi significativement loin de la droite de Henry.

En éliminant X_{13} et X_{11}, la forme du nuage de points (Fig. 7) autorise un ajustement linéaire (par rapport à la droite de Henry).

On remarque aussi que l'écart gaussien (Fig. 8) est trés faible (entre -0.08 et 0.08).

REMARQUE 6. – *En regardant les deux figures 3 et 5, on remarque l'influence de* $X_{13} = -1.40$ *sur la dispersion des autres points par rapport à la droite de Henry. A cause de cela à partir de la figure 3, on n'aperçoit pas l'éloignement de* $X_{11} = 1.01$ *de la droite de Henry. On ne le voit qu'en éliminant* X_{13}.

En utilisant le test de Bol'shev pour le niveau de signification $\alpha = 0.05$, à l'aide du programme Fortran, on prouve que les deux valeurs -1.40 et 1.01 sont aberrantes. On constate donc que le test de Bol'shev détecte simultanément plusieurs

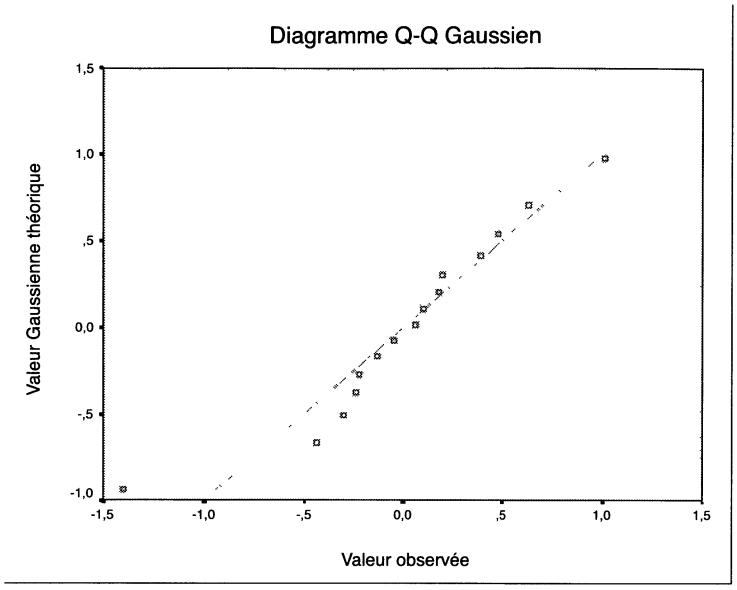

Figure 3

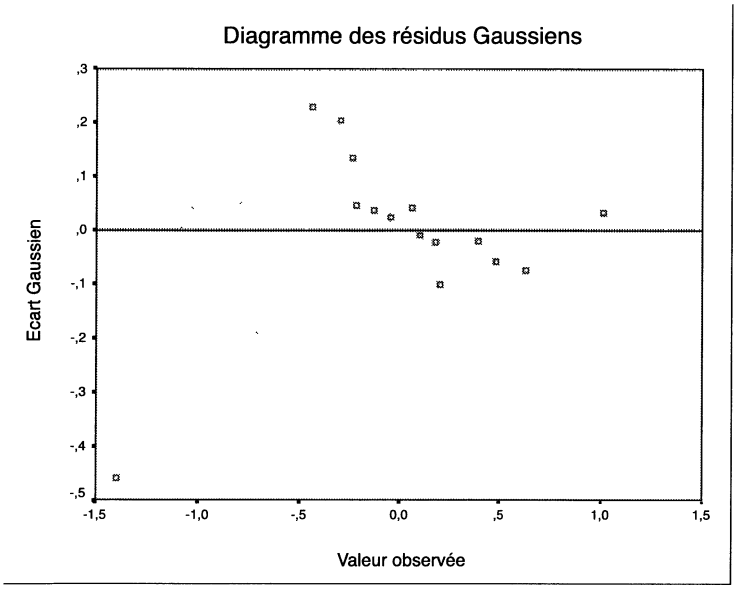

Figure 4

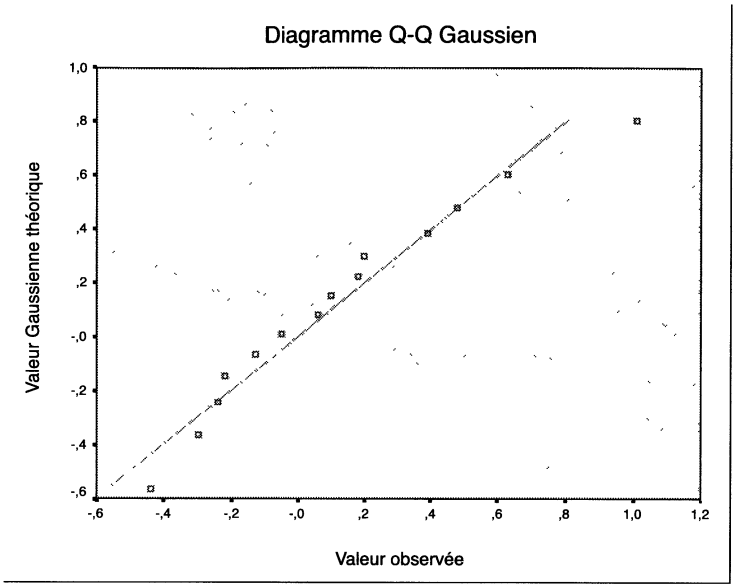

Figure 5

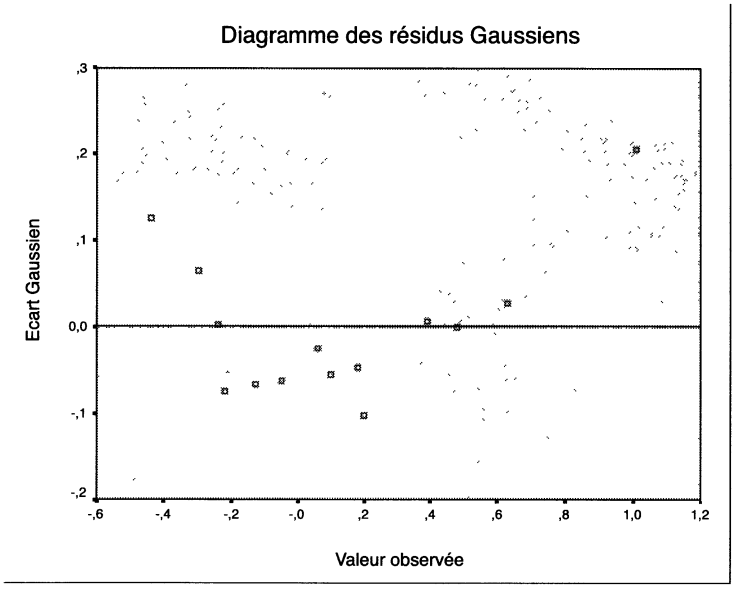

Figure 6

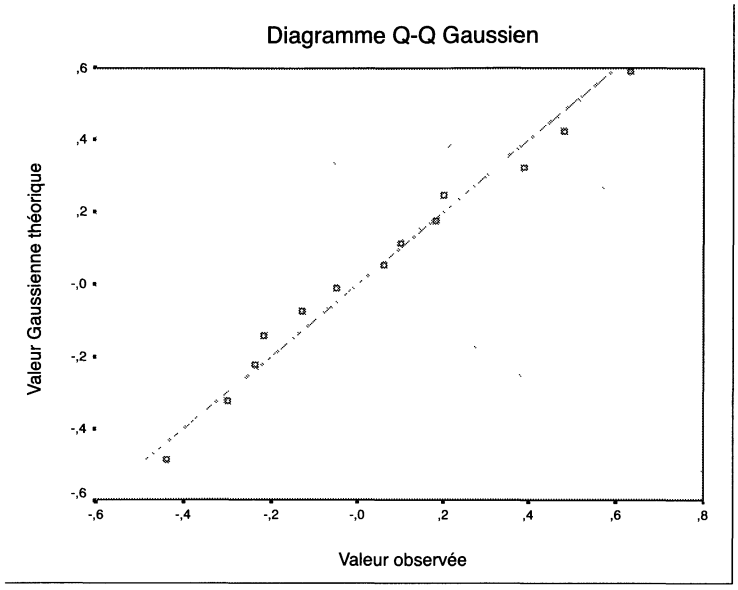

Figure 7

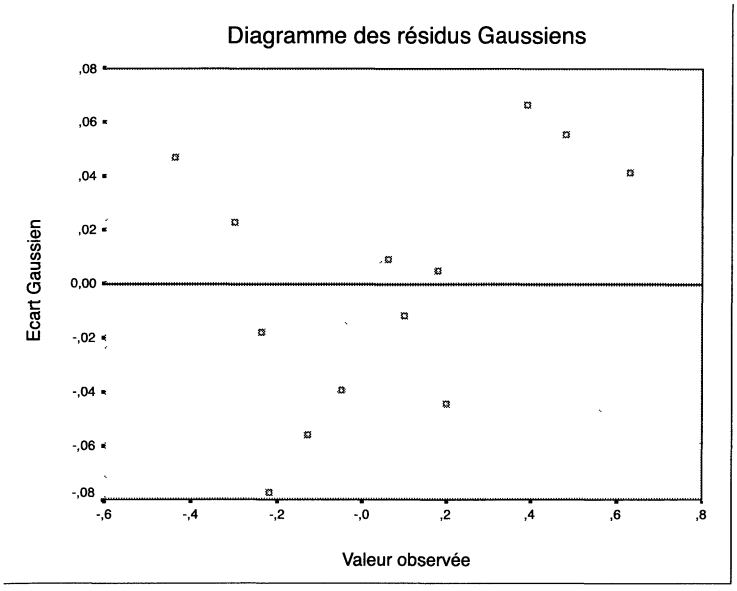

Figure 8

observations aberrantes; on signale que les autres tests comme ceux de "Pearson et Chandra-Sekar" et "Grubbs" ne peuvent pas déceler les valeurs douteuses -1.40 et 1.01 au niveau $\alpha = 0.05$ (voir Barnett et Lewis (1994), page 38).

Références

[1] Vic. BARNETT and Toby LEWIS (1994), "*Outliers in statistical data*", John Wiley and Sons, Inc. : New York.

[2] A. P. BASU (1965), "*On some tests of hypotheses relating to the exponential distribution when some outliers are present*". J. Amer. Statist. Assoc. **60** 548–559. (Reviewer : Benjamin Epstein) 62.25 .

[3] L.N. BOL'SHEV and N.V. SMIRNOV (1965), "*Tables of Mathematical Statistics*", Publishing House "Nauka", pp. 128- 129. (In Russian.)

[4] L. N. BOL'SHEV (1969), "*On tests for rejecting outlying observations*". Trudy In-ta prikladnoi Mat. Tblissi Gosudart.univ.,**2**, pp. 159-177. (In russian.)

[5] L. N. BOL'SHEV and M. UBAIDULLAEVA (1974), "*Chauvenet's test in the classical theory of errors*". Theory Prob. Applications, **19**, pp.683-692.

[6] W. CHAUVENET (1863), "*A Manual of spherical and Practical Astronomy*," II, Philadelphia.

[7] CHIKKAGOUDAR-KUNCHUR (1983), "*Distribution of test statistics for multiple outliers in exponential samples*," Comm. Stat. Theory. and Meth., **12**, 2127-2142.

[8] W. J. DIXON (1950), Analysis of extreme value, *A.M.S.*, **21**, p. 488-506.

[9] W. J. DIXON (1951), Ratios involiving extreme values, *A.M.S.*, **22**, p. 68-78.

[10] P. E. GREENWOOD and M. S. NIKULIN (1996), *A Guide to Chi-Squared Testing*, John Wiley and Sons, Inc. : New York.

[11] F. E. GRUBBS (1950), "*Sample criteria for testing outlying observations*," Ann. Math. Statist., **21**, pp. 27-58.

[12] I.A. IBRAGIMOV and KHALFINA (1978), "*Some asymptotic results concerning the Chauvenet test*," Ter. Veroyatnost. i Primenen., 23, No. **3**, 593-597.

[13] A. G. LAURENT (1963), "*Conditional distribution of order statistics and distribution of the reduced ith order statistic of the exponential model*". Ann. Math. Statist., **34**, pp. 652-657.

[14] M. NIKULIN et A. ZERBET (1999), "*Détection des observations aberrantes par les méthodes statistiques. Partie 1 : Loi normale*," pré-publication No. **9908** du laboratoire de Statistique Mathématique et ses Applications à l'université Victor Segalen Bordeaux 2.

[15] M. NIKULIN (1999), Renyi test, *In : Probability and Mathematical Statistics Encyclopaedia.* (Editor Yu. Prokhorov). Moscou : Big Russian Encyclopaedia, p. 557.

[16] V. I. PAGUROVA (1996), "*On the asymptotic Power of a test for detecting outliers*," Theory Probab. Appl., Vol. 42, No. **3**, pp. 433-443.

[17] E. S. PEARSON and C. CHANDRA SEKAR (1936), "*The efficiency of statistical tools and a criterion for rejection of outlying observation*", Biometrika, **28**, pp. 308-320.

[18] B. PEIRCE (1852), "*Criterion for the rejection of doubtful observations*", Astr. J., **2**, pp. 161-163.

[19] J. N. V. SMIRNOV (1941), "*On a bound for the maximum term in a series of observations* ", Dokl. Akad.Nauk,SSSR(new series), XXXIII, pp. 346-349. (In russian.)

[20] V. G. VOINOV and M. N. NIKULIN (1993), "*Unbiased Estimators and Their Applications*," **Vol.1** : Univariate case, Kluwer Academic Publishers : Dordrecht.

[21] V. G. VOINOV and M. N. NIKULIN (1996), "*Unbiased Estimators and Their Applications*," **Vol.2** : Multivariate Case, Kluwer Academic Publishers : Dordrecht.

[22] S. S. WILKS (1963), "*Multivariate statistical outliers*," Sankhya-ser. -A25, pp. 407-426.

ANNEXE

Démonstration de théorème de Bol'shev-Thompson

On choisit une transformation orthogonale de $\mathbb{R}^n \longrightarrow \mathbb{R}^n$ de la statistique

$$\mathbf{X} = (X_1, ..., X_n)^T \longrightarrow \mathbf{Z} = (Z_1, ..., Z_n)^T$$

de façon que

$$\begin{cases} Z_1 = \sqrt{n}\bar{X}_n \\ Z_{j+1} = \sqrt{\frac{n-j+1}{n-j}} \left(X_j - \frac{1}{n-j+1} \sum_{i=j}^{n} X_i \right) \quad , \quad j = 1, 2, ..., k, \end{cases} \tag{A1}$$

et les autres Z_j, $(j = k+2, ..., n)$ sont construits de façon arbitraire sous la seule condition que la transformation soit orthogonale. De (A1), il résulte que le coefficient de X_j est égal à

$$\sqrt{\frac{n-j+1}{n-j}} - \frac{1}{\sqrt{(n-j)(n-j+1)}} = \sqrt{\frac{n-j}{n-j+1}}$$

alors que les $X_{j+1}, ..., X_n$ ont pour coefficient

$$-\frac{1}{\sqrt{(n-j)(n-j+1)}},$$

on a donc

$$\begin{cases} Z_{j+1} = \sqrt{\frac{n-j}{n-j+1}} X_j - \frac{1}{\sqrt{(n-j)(n-j+1)}} \left(X_{j+1} + ... + X_n \right), \\ Z_1 = \frac{1}{\sqrt{n}} X_1 + \frac{1}{\sqrt{n}} X_2 + ... + \frac{1}{\sqrt{n}} X_n. \end{cases} \tag{A2}$$

Il est alors facile de vérifier que $Z_1, Z_2, ..., Z_{k+1}$ sont centrées, non corrélées et de même variance σ^2, et donc que la transformation effectuée est bien orthogonale.

Il en résulte que :

$$n s_n^2 = \sum_{i=1}^{n} X_i^2 - n \left[\bar{X}_n \right]^2 = \sum_{i=1}^{n} Z_i^2 - Z_1^2 = \sum_{i=2}^{n} Z_i^2. \tag{A3}$$

Posons $\mathbf{V}_k = (V_1, ..., V_k)^T$, où

$$V_i = \frac{1}{s_n \sqrt{n}} Z_{i+1}, \qquad i = 1, 2, ..., k. \tag{A4}$$

De (4), (A2) et (A4) on déduit

$$V_i = \sqrt{\frac{n-i}{n-i+1}}\left(\frac{X_i - \bar{X}_n}{s_n\sqrt{n}} + \frac{\bar{X}_n}{s_n\sqrt{n}}\right) - \frac{1}{\sqrt{(n-i)(n-i+1)}}$$
$$\left(\frac{X_{i+1} - \bar{X}_n}{s_n\sqrt{n}} + ... + \frac{X_n - \bar{X}_n}{s_n\sqrt{n}} + \frac{(n-i)\bar{X}_n}{s_n\sqrt{n}}\right)$$
$$= \sqrt{\frac{n-i+1}{n(n-i)}}\left(Y_i - \frac{1}{n-i+1}(Y_i + Y_{i+1} + ... + Y_n)\right)$$
$$= \sqrt{\frac{n-i+1}{n(n-i)}}\left(Y_i + \frac{1}{n-i+1}(Y_1 + ... + Y_{i-1})\right), \quad i = 1, ..., k, \quad \text{(A5)}$$

puisque $Y_1 + Y_2 + ... + Y_n = 0$.

Comme la distribution de la statistique $\mathbf{V}_k = (V_1, ..., V_k)^T$ est invariante par rapport à toute transformation orthogonale, nous pouvons écrire que la densité $g(v_1, ..., v_k)$ de $\mathbf{V}_k$ est une fonction de $r = \sqrt{v_1^2 + v_2^2 + ... + v_k^2}$:

$$g(v_1, ..., v_k) = f(r),$$

donc (en passant en coordonnées sphériques) on a :

$$\mathbf{P}\{|\mathbf{V}_k| \leqslant R\} = \int ... \int_{0 \leqslant v_1^2 + ... + v_k^2 \leqslant R^2} g(v_1, ..., v_k) dv_1, ..., dv_k$$
$$= \frac{2\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} \int_0^R r^{k-1} f(r) dr, \quad \text{(A6)}$$

où

$$\frac{2\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} r^{k-1}$$

est la surface de la sphère de rayon r dans $\mathbf{R}^k$.

Par ailleurs (A4) entraîne que

$$|\mathbf{V}_k|^2 = \sum_{i=1}^{k} V_i^2 = \sum_{i=1}^{k} \frac{1}{n s_n^2} Z_{i+1}^2 = \frac{Z_2^2 + Z_3^2 + ... + Z_{k+1}^2}{Z_2^2 + Z_3^2 + ... + Z_n^2} \quad \text{(A7)}$$

et donc de (A6) et (A7) on déduit

$$\mathbf{P}\{|\mathbf{V}_k| \leqslant R\} = \mathbf{P}\{|\mathbf{V}_k|^2 \leqslant R^2\} = \mathbf{P}\{\frac{Z_2^2 + Z_3^2 + ... + Z_{k+1}^2}{Z_2^2 + Z_3^2 + ... + Z_n^2} \leqslant R^2\}$$

Etant donné que les Z_i $(2 \leqslant i \leqslant r)$ sont centrés, non corrélés et de même variance σ^2, les Z_i^2/σ^2 suivent indépendamment des lois de χ^2 à un degré de liberté, et donc $Z_i^2/(2\sigma^2) \in \gamma(1/2)$; on en déduit que $\mathbf{V}_k = 2\sigma^2\gamma_{k/2}/(2\sigma^2(\gamma_{k/2}+\gamma_{\frac{n-k-1}{2}}))$ suit une loi beta de paramètres $k/2$ et $\frac{n-k-1}{2}$. On a donc :

$$\mathbf{P}\{|\mathbf{V}_k|^2 \leqslant R^2\} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{n-k-1}{2}\right)} \int_0^{R^2} y^{\frac{k-2}{2}}(1-y)^{(n-k-3)/2}dy$$

$$= I_{R^2}\left(\frac{k}{2}, \frac{n-k-1}{2}\right), \qquad (A8)$$

où

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1}dt, \quad 0 \leqslant x \leqslant 1,$$

est la fonction de répartition de la loi Beta B(a,b) de paramètres a et b.

Par suite,

$$f(r) = \begin{cases} \frac{1}{\pi^{k/2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)}(1-r^2)^{(n-k-3)/2} & , \quad 0 \leqslant r \leqslant 1 \\ 0 \quad , \quad \textit{sinon.} \end{cases} \qquad (A9)$$

Maintenant, en remplaçant les variables $V_1, ..., V_k$ par les variables $Y_1, ..., Y_k$ nous devons faire la transformation (A5) :

$$V_i = \sqrt{\frac{n-i+1}{n(n-i)}}\left(Y_i + \frac{1}{n-i+1}\sum_{j=1}^{i-1} Y_j\right), \; i = 1, 2, ..., k. \qquad (A10)$$

Le Jacobien de cette transformation est :

$$J = \frac{1}{n^{(k-1)/2}(n-k)^{1/2}}, \qquad (A11)$$

d'où le résultat de Bol'shev-Thompson (6).