Documents de travail de la mission biométrie
N° 2-94

Les techniques de rééchantillonnage

Mission biométrie

Centre de coopération internationale
en recherche agronomique pour le développement

Documents de travail de la mission biométrie
N° 2-94

Les techniques de rééchantillonnage

Mission biométrie

Centre de coopération internationale
en recherche agronomique pour le développement

Sommaire

Sommaire 1

Quelques éléments historiques 5
- a Quenouille et la réduction du biais 5
- b Tukey et l'intervalle de confiance des estimateurs : 6
- c Efron et le Bootstrap : 6
- d La validation croisée 7

Le Jackknife : réduction du biais 9
- 1.1 Le biais de quelques estimateurs 9
  - a l'estimateur de la variance 9
  - b l'estimateur de l'écart-type 10
  - c l'estimateur de la borne d'une loi uniforme 11
  - d l'estimateur du coefficient de corrélation 11
  - e l'estimateur d'un quotient 11
  - f conclusion sur le biais 12
- 1.2 L'estimateur du Jackknife 12
- 1.3 Des exemples d'estimateurs 14
  - a l'estimateur de la variance 14
  - b l'estimateur de l'écart-type 14
  - c l'estimateur de la borne d'une loi uniforme 15
  - d l'estimateur du coefficient de corrélation 17
  - e l'estimateur d'un quotient 17
- 1.4 Le Jackknife et le biais 19
  - a le Jackknife et les estimateurs sans biais 19
  - b estimation du biais 19
  - c ce n'est pas l'estimateur du moindre biais 20
- 1.5 Variantes sur le Jackknife 20
  - a élimination des termes en $1/n^2$ 20
  - b le Jackknife de groupe 21

Variance du Jackknife 25
- 2.1 Variance de l'estimateur du Jackknife 25
- 2.2 Démarche suivie par Efron 28
- 2.3 Exemples de variances d'estimateurs 29
  - a l'estimateur de la moyenne 29
  - b l'estimateur fonction de la moyenne 30
  - c l'estimateur de la variance 31

d l'estimateur de l'écart-type 32
e l'estimateur d'un quotient 33
f méthode de "capture-recapture" 34
2.4 Contre-exemple des spéculations de Tuckey 37

Le Bootstrap **41**
3.1 Introduction 41
a Substitution de la distribution empirique à la distribution théorique 41
b Utilisation de la simulation 42
c La méthode du Bootstrap 43
3.2 Exemples 43
a La médiane 43
b Variance de l'estimateur du coefficient de variation 46
3.3 Taille de l'échantillon du Bootstrap 47
3.4 Comparaison Jackknife - Bootstrap 48

Application à l'étude d'une régression non linéaire **53**
a Présentation du problème 53
b Modèle utilisé 54
c Estimation des paramètres par les méthodes de rééchantillonnage 57
d Conclusion 60

Etude de la stabilité de l'ACP par le Bootstrap **61**
5.1 Introduction 61
5.2 Rappels 61
a But de l'ACP 61
b Principe de l'ACP 61
5.3 Application du Bootstrap à l'ACP 62
a Principe du Bootstrap 62
b Tirage des sous-échantillons 62
c Notations 63
5.4 Evaluation de la stabilité de l'ACP 64
a Etude de la stabilité de l'ACP non-normée 64
b Etude de la stabilité de l'ACP normée 65
5.5 Mesures de la stabilité de l'ACP 66
a Stabilité des valeurs propres 67
b Stabilité des sous-espaces de représentation 67
5.6 Exemples 68
a Composition du lait 69
b Accroissements du poids des chèvres 70
c Conclusions générales 72

La validation croisée **75**
6.1 Présentation de la validation croisée 75
a Définition 75
b Justification intuitive 75
c Domaines d'utilisation 76
d Variantes de la validation croisée 76
6.2 Application à l'interpolation par Krigeage 77
a Présentation sommaire du krigeage 77
b Objet et mise en oeuvre de la validation croisée 78
c Résultats 79
6.3 Conclusion 80

Annexes **93**
A–1 La démonstration d'Efron 93

Quelques éléments historiques

La plupart des approches statistiques reposent sur des hypothèses distributionnelles et, en général, sur l'hypothèse de normalité. Ce postulat est, dans bien des cas, raisonnable. Les propriétés qui lui sont liées permettent des développements théoriques intéressants conduisant à des outils efficaces et qui ne nécessitent pas des quantités de calculs irréalisables.

Cependant, il se trouve des situations où la normalité n'est pas acceptable. Soit on connaît la réelle loi de distribution; si ses propriétés sont intéressantes, des développements sont encore possibles mais ce n'est pas toujours le cas. Soit on sait rejeter la normalité mais sans connaître la forme de la distribution. La recherche de méthodes alternatives a conduit, en reprenant les termes d'Efron (1979), à "penser l'impensable", l'impensable étant "d'effectuer 500.000 opérations dans l'analyse d'observations faites sur 16 individus". L'idée est donc, quand cela est possible, de remplacer certaines hypothèses par la répétition d'un calcul sur un grand nombre de sous-échantillons.

a Quenouille et la réduction du biais

Historiquement, on peut attribuer les premiers pas de cette approche à Quenouille. Il s'intéressait à la réduction du biais dans l'estimation du coefficient de corrélation et proposa en 1949 de découper l'échantillon de travail en deux sous-échantillons.

Soit θ un paramètre à estimer, le coefficient de corrélation par exemple, $\hat{\theta}$ son estimation sur l'échantillon total, $\hat{\theta}_1$ et $\hat{\theta}_2$ les estimations sur les deux sous-échantillons

Quenouille pose les bases du Jackknife lorsqu'il montre que l'estimateur $\hat{\theta}^* = 2\hat{\theta} - \frac{1}{2}\left(\hat{\theta}_1 + \hat{\theta}_2\right)$ est de biais inférieur à $\hat{\theta}$.

En 1956 - l'idée a mis sept ans pour mûrir - Quenouille généralise le principe en découpant l'échantillon (de taille n) en g sous-échantillons de taille h ($n = gh$). $\hat{\theta}_{(-i)}$ est l'estimateur obtenu sur l'ensemble des individus en omettant le $i^{ème}$ échantillon.

On définit alors les pseudo-valeurs (terme introduit par Tukey en 1962)

$$\tilde{\theta}_{(i)} = g\hat{\theta} - (g-1)\hat{\theta}_{(-i)}$$

L'estimateur du Jackknife est alors la moyenne de ces pseudo-valeurs :

$$\tilde{\theta} = \frac{1}{g}\sum_{i=1}^{g}\tilde{\theta}_{(i)} = g\hat{\theta} - \frac{g-1}{g}\sum_{i=1}^{g}\hat{\theta}_{(-i)}$$

Si g vaut deux, on retrouve l'estimateur de 1949.

b Tukey et l'intervalle de confiance des estimateurs :

L'objectif de Quenouille est uniquement la réduction du biais des estimateurs. Quelques années plus tard, Tukey (1958) étend l'idée à l'estimation des intervalles de confiance de ces estimateurs. (Il est l'auteur du terme Jackknife = couteau Suisse au sens d'outil à tout faire, d'où le terme proposé en français d'Eustache).

Il conjecture que

$$\frac{\tilde{\theta} - \theta}{s}$$

suit une distribution de Student à $g-1$ degré de liberté d'où le passage à l'intervalle de confiance. (s est l'écart-type de $\tilde{\theta}$).

Cette conjecture sera ensuite montrée vraie dans un grand nombre de cas, au moins asymptotiquement et donc lorsque g est grand. Il en découlera la forme habituelle du Jackknife où g est maximisé à n, h étant alors de 1. Ceci revient à considérer n sous-échantillons de $n-1$ individus en omettant tour à tour chacun des individus.

c Efron et le Bootstrap :

On peut considérer que chaque estimateur $\hat{\theta}_i$ est une fonction des observations X_i et de poids P_i associés à chaque observation.

$$\hat{\theta}_i = T\left(X_1, \ldots, X_i, \ldots, X_n, P_1, \ldots, P_i, \ldots, P_n\right)$$

Dans le Jackknife classique, le poids du $i^{ème}$ individu est mis à zéro et les autres à

$$\frac{1}{n-1} \qquad \left(\sum P_i = 1\right)$$

$$\hat{\theta}_i = T\left(X_1, \ldots, X_i, \ldots, X_n, \frac{1}{n-1}, \ldots, 0, \frac{1}{n-1}\right)$$

Efron en 1979, 20 ans après Tukey, introduit la méthode du Bootstrap. Le terme de Bootstrap fait référence au tirant de botte avec l'idée d'essayer de se soulever en tirant sur ses lacets. Le terme Cyrano est proposé en français.

Les sous-échantillons sont tirés avec remise dans l'ensemble des observations $(X_1, \ldots, X_i, \ldots, X_n)$, certaines seront donc représentées plusieurs fois alors que d'autres seront absentes.

$$\hat{\theta}_B = T\left(X_1, \ldots, X_i, \ldots, X_n, P_1, \ldots, P_i, \ldots, P_n\right) \text{ où } P_i \in \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$$

Le poids d'un individu sera alors sa fréquence de tirage

Le Bootstrap revient donc à considérer que l'on a un très grand nombre d'échantillons alors que l'on n'en a qu'un seul. Il est clair que l'on suppose que l'échantillon de base est représentatif sinon tous les rééchantillons seront aussi biaisés que le groupe initial et les résultats sans intérêt.

De nombreuses applications ont été publiées : en régression, en analyse multidimensionnelle, et dans un grand nombre d'études particulières : taxonomie numérique, cartographie du génome, etc...

d La validation croisée

Parmi ces méthodes de rééchantillonnages, on doit inclure le principe de validation croisée. L'idée est ancienne et a certainement une paternité multiple. Elle s'est développée avec l'accroissement des moyens informatiques.

En régression, par exemple, on estime souvent les paramètres du modèle et on teste la pertinence de ce modèle sur les mêmes données, d'où une vision évidemment optimiste. L'idée est ici de séparer aléatoirement l'échantillon en deux parties, l'une servant à établir le modèle, l'autre à le tester. Une approche identique est souvent utilisée en analyse discriminante.

Il a ensuite été montré qu'il n'est pas nécessaire de partager l'échantillon en deux parties égales. Il est aujourd'hui plus commun d'enlever un seul point à chaque fois et de voir comment le modèle ajuste le point exclu. L'analogie avec le Jackknife est évidente mais sans l'idée de réduction de biais ou d'intervalle de confiance.

Nous avons donc là un ensemble de méthodes non paramétriques, que l'on considère, en partie abusivement, libres de contraintes de la théorie paramétrique. L'intérêt est évident pour les modèles où les solutions théoriques ne sont pas disponibles. Cependant des limites existent et comme conclut Efron, la compréhension de ces limites pourrait mettre en valeur les vertus de la théorie paramétrique et peut-être suggérer des compromis efficaces.

Ce document se propose d'expliciter ces diverses approches et de montrer des exemples d'applications à des méthodes comme l'ACP ou la régression.

Le Jackknife : réduction du biais

Soit X une variable aléatoire de fonction de répartition $\mathcal{F}$. Nous pouvons lui associer la suite $(X_1, X_2, \ldots, X_n)$ de n variables aléatoires indépendantes de même loi que X (nous parlerons d'un n-échantillon issu de la variable parente X).

La réalisation de l'échantillon $(X_1, X_2, \ldots, X_n)$ sera notée $(x_1, x_2, \ldots, x_n)$.

Généralement la loi de X sera fonction d'un paramètre θ que nous chercherons à estimer, au moyen d'un estimateur :

$$T(X_1, X_2, \ldots, X_n)$$

la valeur prise par cet estimateur étant l'estimation de θ :

$$\hat{\theta} = T(x_1, x_2, \ldots, x_n)$$

Remarque : Nous parlerons aussi, pour désigner T, de statistique.

Si nous désignons par $E\left[T\left(X_1, X_2, \ldots, X_n\right)\right]$, l'espérance de la statistique T, (lorsqu'elle existe), par rapport à la loi $\mathcal{F}$, la valeur de l'écart :

$$E\left[T\left(X_1, X_2, \ldots, X_n\right)\right] - \theta$$

est le biais de T.

Nous noterons parfois T_n l'estimateur $T(X_1, X_2, \ldots, X_n)$ et $T_{(-i)} = T(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ l'estimateur calculé à partir du sous-échantillon de taille $(n-1)$ $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.

1.1 Le biais de quelques estimateurs

Dans de nombreux cas, nous pouvons montrer que le biais d'un estimateur T_n est de la forme :

$$E(T_n) - \theta = b_n = \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots$$

Les a_i étant indépendants de n.

a l'estimateur de la variance

Nous considérons un échantillon de n variables aléatoires identiques et indépendantes $(X_1, X_2, \ldots, X_n)$ avec :

$$\forall i, \qquad X_i \sim \mathcal{N}(\mu, \gamma)$$

L'estimateur du maximum de vraisemblance de la variance γ s'écrit :

$$V_n = \frac{1}{n} \sum \left(X_i - \bar{X}\right)^2$$

Cet estimateur est biaisé :

$$E\left(V_n\right) = \gamma - \frac{1}{n}\gamma \quad \Rightarrow \quad b_n = -\frac{1}{n}\gamma$$

D'où : $a_1 = -\gamma \quad a_2 = 0 \cdots$

b l'estimateur de l'écart-type

Nous considérons un échantillon de n variables aléatoires identiques et indépendantes $(X_1, X_2, \ldots, X_n)$ avec :

$$\forall i, \qquad X_i \sim \mathcal{N}(\mu, \gamma)$$

Nous savons que l'estimateur du maximum de vraisemblance de $\gamma = \sigma^2$ s'écrit :

$$\frac{1}{n} \sum \left(X_i - \bar{X}\right)^2$$

qu'il est biaisé et que nous utilisons classiquement :

$$S^2 = \frac{1}{n-1} \sum \left(X_i - \bar{X}\right)^2$$

avec :

$$E\left(S^2\right) = \sigma^2 \quad \text{et} \quad \frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

Pour estimer σ, nous utilisons généralement la statistique S mais son espérance $E(S)$ est différente de σ puisque :

$$E\left(\sqrt{S^2}\right) \neq \sqrt{E\left(S^2\right)}$$

Nous pouvons montrer que :

$$E(S) = \sigma c_n \quad \text{avec} \quad c_n = \left(\frac{2}{n-1}\right)^{1/2} \frac{\Gamma\left(\frac{1}{2}n\right)}{\Gamma\left(\frac{1}{2}n - \frac{1}{2}\right)}$$

Rappel : $\Gamma(u) = \int_0^\infty t^{u-1} \exp(-t) dt$

Quelques valeurs de c_n :

n	4	6	8	10	12	16	20	30	50	100
c_n	0,9213	0,9515	0,9650	0,9727	0,9776	0,9835	0,9869	0,9914	0,9949	0,9975

c l'estimateur de la borne d'une loi uniforme

Soit X distribuée selon une loi uniforme sur $[0, \theta]$ et $(x_1, \ldots, x_n)$ une réalisation de l'échantillon $(X_1, X_2, \ldots, X_n)$. L'estimateur du maximum de vraisemblance de θ est :

$$S_n = \operatorname*{Sup}_i X_i$$

Nous pouvons montrer alors :

$$E(S_n) - \theta = -\frac{\theta}{n+1} = \frac{\theta}{n} - \frac{\theta}{n^2} + \frac{\theta}{n^3} - \cdots$$

D'où : $a_1 = \theta \quad a_2 = -\theta \cdots$

d l'estimateur du coefficient de corrélation

Soit $(X_1, Y_1) \ldots (X_i, Y_i) \ldots (X_n, Y_n)$ un n-échantillon de couples de variables aléatoires normales. L'estimateur du coefficient de corrélation ρ est égal à :

$$R = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum \left(X_i - \bar{X}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2}}$$

Si nous notons (Kendall - Stuart [10]) :

$$F(\alpha, \beta, \gamma, t) = 1 + \frac{\alpha\beta}{\gamma}\frac{t}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{t^2}{2!} + \cdots$$

nous pouvons montrer :

$$\begin{aligned} E(R) &= \frac{\rho\Gamma^2\left(\frac{1}{2}n\right)}{\Gamma\left(\frac{1}{2}(n-1)\right)\Gamma\left(\frac{1}{2}(n+1)\right)} F\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}(n+1), \rho^2\right) \\ &= \rho\left(1 - \frac{1-\rho^2}{2n} + o\left(n^{-2}\right)\right) \end{aligned}$$

Rappel : $\begin{cases} \Gamma(u) = \int_0^\infty t^{u-1}\exp(-t)dt \\ o\left(n^{-2}\right) \text{ désigne une fonction telle que } n^2 o\left(n^{-2}\right) \to 1 \text{ quand } n \to \infty \end{cases}$

e l'estimateur d'un quotient

Soit X et Y deux variables aléatoires, auxquelles nous associons les suites $(X_1, X_2, \ldots X_n)$ et $(Y_1, Y_2, \ldots Y_n)$. Nous poserons $\mu = E(X_i)$ et $\eta = E(Y_i)$. Le problème est de vouloir estimer le rapport $\theta = \frac{\mu}{\eta}$, car l'estimateur $R = \frac{\sum Y_i}{\sum X_i} = \frac{\bar{Y}}{\bar{X}}$, est un estimateur biaisé. Nous pouvons montrer que :

$$E(R) = \theta - \frac{\mathrm{Cov}(R, \bar{X})}{E(\bar{X})}.$$

Certaines études ont été faites pour évaluer $E(R)$ dans différents cas. Par exemple Durbin [3] étudie la forme de $E(R)$ lorsque la régression entre X et Y est linéaire et X distribuée de façon normale. Il montre que, dans un tel cas :

$$E(R) - \theta = a\left(\frac{1}{n} + \frac{3}{n^2} + \frac{15}{n^3}\right) + o\left(n^{-4}\right)$$

De nombreuses personnes ont proposé d'autres estimateurs pour estimer un rapport [11], et même si le Jackknife n'est pas toujours le meilleur estimateur, il s'en approche beaucoup.

f conclusion sur le biais

Il est bien évident que tous les estimateurs n'ont pas un biais qui peut s'exprimer sous cette forme.

Le problème de la structure de T ne semble pas complètement éclairci.

1.2 L'estimateur du Jackknife

L'idée de base du Jackknife provient du fait que la donnée des $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ donc d'un n-échantillon, est la donnée de plusieurs $(n-r)$-échantillons.

Intuitivement, cela devrait apporter une information sur les biais : $b_1, b_2, \ldots, b_n$ (b_i est le biais d'un échantillon de taille i).

Il y a donc peut-être moyen de calculer ainsi certains des paramètres $a_1, a_2, \ldots$ du développement :

$$b_n = \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots$$

Les a_i étant indépendants de n.

Notations :

T	l'estimateur $T_n(X_1, X_2, \ldots\ldots X_n)$
$T_{(-i)}$	l'estimateur $T_{n-1}(X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots X_n)$
$T_{(.)}$	$= \frac{1}{n}\sum_{i=1}^{n} T_{(-i)}$
$\hat{\theta}$	la valeur prise par $T_n(x_1, x_2, \ldots x_n)$
$\hat{\theta}_{(-i)}$	la valeur prise par $T_{n-1}(x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots x_n)$

Considérons les n estimateurs de θ construits à partir des $(n-1)$-échantillons :

$$E\left[T\left(X_1, X_2, \cdots X_{i-1}, X_{i+1}, \cdots X_n\right)\right] = \theta + b_{n-1}$$

Nous avons : $\begin{cases} E(T) & = \theta + \dfrac{a_1}{n} + \dfrac{a_2}{n^2} + \cdots \\ E\left(T_{(.)}\right) & = \theta + \dfrac{a_1}{n-1} + \dfrac{a_2}{(n-1)^2} + \cdots \end{cases}$

L'estimateur du Jackknife $\tilde{T}$ est une combinaison linéaire de T et $T_{(.)}$.

$$\boxed{\tilde{T} = nT - (n-1)T_{(.)}}$$

$$\begin{aligned} E(\tilde{T}) &= n\theta + a_1 + \frac{a_2}{n} + \frac{a_3}{n^2} + \cdots - \left((n-1)\theta + a_1 + \frac{a_2}{n-1} + \cdots\right) \\ &= \theta + a_2\left(\frac{1}{n} - \frac{1}{n-1}\right) + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \cdots \\ &= \theta - \frac{a_2}{n(n-1)} + \frac{1-2n}{n^2(n-1)^2}a_3 + \cdots \\ &= \theta + o\left(n^{-2}\right) \end{aligned}$$

<u>Conclusion</u> : Asymptotiquement, le biais de $\tilde{T}$ est moindre que celui de T .

Remarque :

$$\begin{aligned} \tilde{T} &= nT - (n-1)T_{(.)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[T - (n-1)\left(T_{(-i)} - T\right)\right] \end{aligned}$$

Les valeurs $T - (n-1)\left(T_{(-i)} - T\right)$ lorsque i varie de 1 à n, sont les **"pseudo-valeurs"** introduites par Tukey [15]. Elles sont notées $\tilde{T}_{(i)}$.

$$\boxed{\tilde{T} = \frac{1}{n}\sum_{i=1}^{n}\tilde{T}_{(i)}}$$

Nous noterons les valeurs prises par ces "pseudo-valeurs" :

$$\tilde{\theta}_{(i)} = \hat{\theta} - (n-1)\left(\hat{\theta}_{(-i)} - \hat{\theta}\right) = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$$

et $\tilde{\theta}$ la valeur prise par $\tilde{T}$ l'estimateur du Jackknife qui est la moyenne de ces "pseudo-valeurs".

1.3 Des exemples d'estimateurs

a l'estimateur de la variance

Soit la réalisation d'un échantillon :

i	1	2	3	4	5
x_i	$-1,0$	$0,0$	$0,5$	$1,0$	$4,0$

la valeur prise par l'estimateur de la variance est :

$$\hat{\gamma} = \frac{1}{n}\sum (x_i - \bar{x})^2 = 2,84$$

les valeurs prises par cet estimateur sur un 4-échantillon :

$\hat{\gamma}_{(-i)}$	2,42	3,29	3,50	3,55	0,54

Calculons les "pseudo-valeurs" : $\tilde{\gamma}_{(i)} = 5\hat{\gamma} - 4\hat{\gamma}_{(-i)}$

$\tilde{\gamma}_{(i)}$	4,52	1,04	0,20	0,00	12,04

La moyenne de ces "pseudo-valeurs" est l'estimateur du Jackknife :

$$\tilde{\gamma} = \frac{1}{5}\sum_{i=1}^{5}\tilde{\gamma}_{(i)} = 3,56$$

b l'estimateur de l'écart-type

Pour étudier l'estimateur du Jackknife, nous allons procéder par simulation des variables :

$$\forall i, \quad X_i \sim \mathcal{N}(0,10) \quad \text{et} \quad \sigma = \sqrt{10} = 3,1623$$

Soit un échantillon $(x_1, x_2, \ldots, x_n)$, nous calculons :

$$\hat{s}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 \quad , \quad \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \quad \text{et} \quad \hat{s}$$

$$\hat{s}^2_{(-i)} = \frac{1}{n-2}\sum_{j\neq i}\left(x_j - \bar{x}_{(-i)}\right)^2 \quad , \quad \bar{x}_{(-i)} = \frac{1}{n-1}\sum_{j\neq i} x_j \quad \text{et} \quad \hat{s}_{(-i)}$$

les "pseudo-valeurs" : $\tilde{s}_{(i)} = n\hat{s} - (n-1)\hat{s}_{(-i)}$

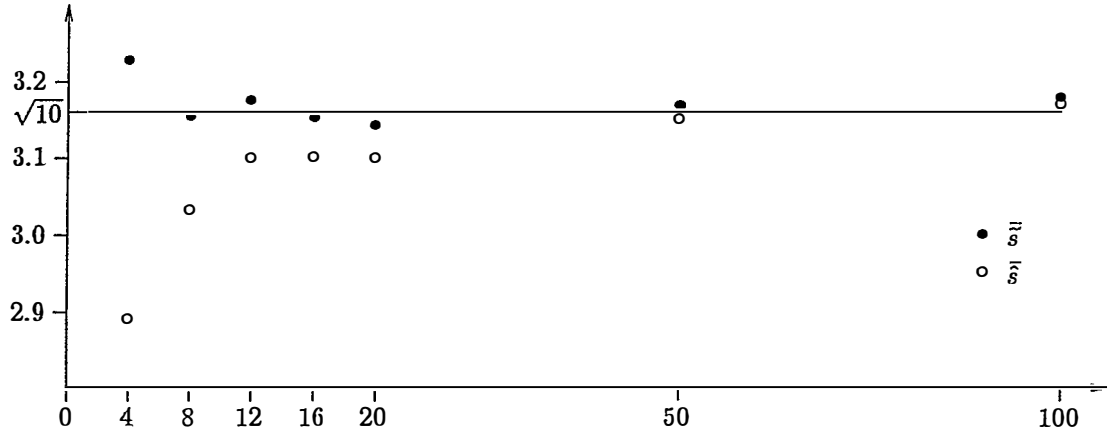
et l'estimateur du Jackknife : $\tilde{s} = \frac{1}{n}\sum_{i=1}^{n}\tilde{s}_{(i)}$

Nous répétons l'opération 500 fois et nous obtenons les distributions empiriques de $\hat{s}$ et $\tilde{s}$:

$$\begin{array}{ccccc} \hat{s}_1 & \hat{s}_2 & \hat{s}_3 & \dots & \hat{s}_{500} \\ \tilde{s}_1 & \tilde{s}_2 & \tilde{s}_3 & \dots & \tilde{s}_{500} \end{array}$$

Moyenne des $\hat{s}_i$ et des $\tilde{s}_i$ pour différentes valeurs de n ($\sigma = 3,1623$) :

n	4	8	12	16	20	50	100
$\hat{s}$	2,8896	3,0315	3,0994	3,0990	3,0984	3,1505	3,1717
$c_n\sigma$	2,9134	3,0516	3,0914	3,1101	3,1209	3,1462	3,1543
$\bar{\tilde{s}}$	3,2283	3,1541	3,1744	3,1524	3,1414	3,1667	3,1798
$\lvert\hat{s} - \sigma\rvert$	0,2727	0,1107	0,0633	0,0633	0,0639	0,0118	0,0094
$\lvert\bar{\tilde{s}} - \sigma\rvert$	0,0660	0,0082	0,0121	0,0099	0,0209	0,0044	0,0175



c l'estimateur de la borne d'une loi uniforme

Soit X une variable aléatoire distribuée de façon uniforme sur $[0, \theta]$ et le n -échantillon $(X_1, X_2, \dots X_n)$ issu de la variable parente X.

Nous utiliserons la notation suivante : $X_{(i)}$ désignant la variable X_j de rang i :

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

L'estimateur du maximum de vraisemblance de θ est :

$$\hat{\theta} = \operatorname*{Sup}_i X_i = X_{(n)}$$

En ôtant la variable X_i, cet estimateur devient :

$$\hat{\theta}_{(-i)} = \begin{cases} X_{(n)} & \text{si } i \neq n \\ X_{(n-1)} & \text{si } i = n \end{cases}$$

Les pseudo-valeurs sont égales à :

$$\tilde{\theta}_{(i)} = \begin{cases} nX_{(n)} & \text{si } i \neq n \\ nX_{(n)} - (n-1)X_{(n-1)} & \text{si } i = n \end{cases}$$

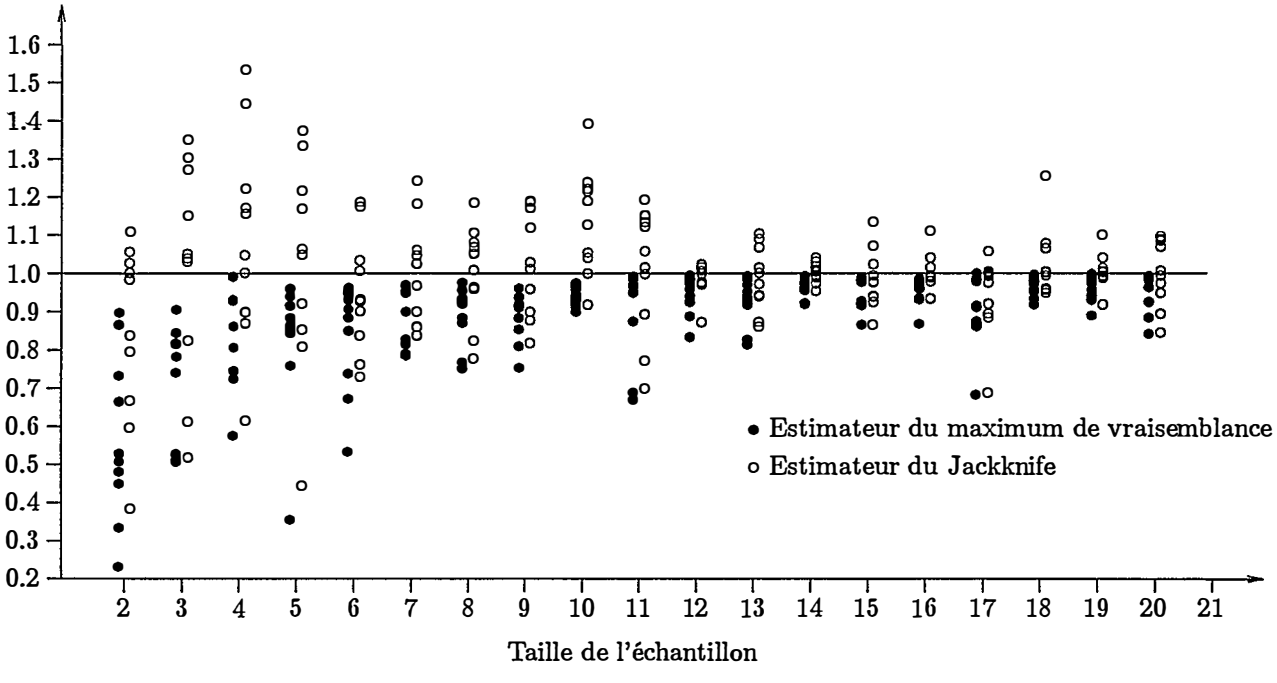
L'estimateur du Jackknife s'écrit :

$$\tilde{\theta} = \sum_{i=1}^{n} \tilde{\theta}_{(i)} = X_{(n)} + \frac{n-1}{n}\left(X_{(n)} - X_{(n-1)}\right)$$

et la variance de cet estimateur est égale à :

$$\mathrm{Var}\left(\tilde{\theta}\right) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2$$

Dans notre exemple, nous prendrons $\theta = 1$. Nous tirerons aléatoirement une dizaine d'échantillons de taille n et nous ferons varier n de 2 à 20.
Pour chacun de ces échantillons, nous calculerons la valeur prise par l'estimateur du maximum de vraisemblance $\hat{\theta} = \operatorname{Sup}_i x_i$, et par l'estimateur du Jackknife $\tilde{\theta}$.
Les valeurs obtenues sont présentées sur un graphe dont les abscisses correspondent à la taille de l'échantillon.



Nous remarquons :

- Les valeurs prises par l'estimateur sont toujours plus faibles que la valeur de la borne. La différence, ou le biais, diminuant lorsque n augmente.

- Les valeurs prises par l'estimateur du Jackknife semblent moins biaisées que l'estimateur du maximum de vraisemblance, mais celles-ci sont, par contre, plus étalées.

d l'estimateur du coefficient de corrélation

Etudions la relation entre le taux de goudron (X) et le taux de nicotine (Y) [1]. L'échantillon mesuré porte sur 10 marques de cigarettes :

x_i	0,45	0,77	1,07	1,03	1,34	1,14	1,15	0,90	0,55	1,15
y_i	11,0	13,0	14,0	15,0	17,0	18,0	14,5	13,5	8,5	16,5

La valeur prise par l'estimateur du coefficient de corrélation est $\hat{\rho} = 0,89$.

$\hat{\rho}_{(-i)}$	0,90	0,89	0,90	0,89	0,88	0,91	0,91	0,89	0,87	0,88
$\tilde{\rho}_{(i)}$	0,79	0,89	0,79	0,89	0,98	0,71	0,71	0,89	1,07	0,98

La valeur prise par l'estimateur du Jackknife est $\tilde{\rho} = \frac{1}{n}\sum \tilde{\rho}_{(i)} = 0,87$

e l'estimateur d'un quotient

Etudions le taux de surface boisée en France. Les données recueillies proviennent de l'inventaire forestier national de 1983 [2], et correspondent aux données de 90 départements[2]. Nous connaissons la surface de la France, qui est de 54.918.450 ha[3]. Nous connaissons aussi exactement la superficie totale boisée, qui est de 13.775.471 ha, d'où le taux de boisement national de 25,0835 %.

Oublions les données précédentes. Nous connaissons toujours la surface de la France, mais nous voulons estimer la superficie nationale boisée ou, ce qui revient au même, le taux de boisement national. Pour cela nous utiliserons l'estimateur du quotient $Q = \frac{\sum Y_i}{\sum X_i} = \frac{\bar{Y}}{\bar{X}}$. Notre échantillon sera composé de quelques départements français.

Soit 10 départements tirés au hasard :

[1] Selon la loi n° 91.32 "Nuit gravement à la santé"

[2] la Corse et l'Île-de-France sont comptées chacune pour deux départements

[3] Cette superficie diffère de la superficie nationale officielle de 54.919.193 ha, car les données de surface de chaque département sont celles des différentes années d'inventaire

Code	Département	Surface boisée	Superficie du département	Taux de boisement
01	Ain	177.071	578.501	30, 61
12	Aveyron	224.603	877.122	25, 61
17	Charente Maritime	101.706	690.000	14, 74
34	Hérault	138.484	622.673	22, 24
37	Indre et Loire	135.776	615.403	22, 06
44	Loire Atlantique	42.515	695.640	6, 11
54	Meurthe et Moselle	169.201	527.737	32, 06
74	Haute – Savoie	170.814	483.862	35, 30
83	Var	280.016	603.250	46, 42
87	Haute – Vienne	135.210	555.825	24, 33
Total		1.575.396	6.250.013	

Calculons la valeur prise par l'estimateur du quotient :

$$\hat{q} = \frac{1.575.396}{6.250.013} * 100\ \% = 25,206\ \%$$

Calculons cet estimateur en omettant un par un les départements, et calculons les "pseudo-valeurs" :

Code	$\hat{q}_{(-i)}$	$\tilde{q}_{(i)}$
01	24, 655	30, 166
12	25, 141	25, 794
17	26, 505	13, 516
34	25, 534	22, 252
37	25, 550	22, 116
44	27, 598	3, 683
54	24, 574	30, 896
74	24, 359	32, 831
83	22, 940	45, 601
87	25, 292	24, 433
Moyenne		25, 129

La valeur prise par l'estimateur du Jackknife est :

$$\tilde{q} = \frac{1}{10}\sum_{i=1}^{10} \tilde{q}_{(i)} = 25,129\ \%$$

Ce résultat est à comparer avec un autre estimateur $R = \frac{1}{n}\sum \frac{Y_i}{X_i}$, qui représente le taux moyen de boisement par département.
La valeur prise par cet estimateur est : $\hat{r} = 25,948\ \%$.

1.4 Le Jackknife et le biais

a le Jackknife et les estimateurs sans biais

Supposons que l'estimateur du Jackknife $\tilde{T}$ prenne les mêmes valeurs que l'estimateur T $\left(T \equiv \tilde{T}\right)$, nous avons :

$$\begin{aligned} T = \tilde{T} &= nT - (n-1)T_{(.)} \\ T &= T_{(.)} \\ E(T) &= E\left(T_{(.)}\right) \end{aligned}$$

Nous savons, dans ce cas, que l'espérance d'un estimateur construit sur un n-échantillon est égale à l'espérance d'un estimateur construit sur un $(n-1)$-échantillon.

$$E\left(T_n\right) = E\left(T_{(.)}\right) = E\left(T_{n-1}\right)$$

<u>Conclusion :</u>

Si $\tilde{T} \equiv T$ quel que soit n, $E\left(T_n\right) = E\left(T_m\right)$ quels que soient n et m.
Si l'estimateur est convergent : il est sans biais.
Si $E(T) = \theta$, il est évident que $E\left(\tilde{T}\right) = \theta$.

b estimation du biais

Quenouille [14] propose d'utiliser $(n-1)\left(T_{(.)} - T\right)$ comme estimateur du biais de T.

$$\tilde{T} = nT - (n-1)T_{(.)} = T - \underbrace{(n-1)\left(T_{(.)} - T\right)}_{\text{correction apportée à } T}$$

Nous avons :

$$E\left(T_{(.)} - T\right) = \underbrace{\frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \ldots}_{\text{biais } b_{n-1}} - \underbrace{\frac{a_1}{n} + \frac{a_2}{n^2} + \ldots}_{\text{biais } b_n}$$

Si nous notons B_n l'estimateur du biais b_n :

$$\begin{aligned} E\left(B_n\right) &= (n-1)\left(b_{n-1} - b_n\right) \\ E\left(B_n\right) - b_n &= (n-1)b_{n-1} - nb_n \end{aligned}$$

Pour que ce biais soit nul, il faut que $b_n = \dfrac{n-1}{n} b_{n-1}$ et pour que ceci soit vérifié, pour tout n il faut que b_n soit une fonction du type $\dfrac{a_1}{n}$.

c ce n'est pas l'estimateur du moindre biais

L'estimateur du Jackknife est une combinaison linéaire de T et $T_{(.)}$.

$$\tilde{T} = nT - (n-1)T_{(.)}$$

mais "a priori" ce n'est pas l'estimateur du moindre biais que l'on peut construire parmi les combinaisons linéaires de T et $T_{(.)}$.

$$\begin{aligned} E(T) &= \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \cdots \\ E\left(T_{(.)}\right) &= \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \frac{a_3}{(n-1)^3} + \cdots \end{aligned}$$

Remarque :

$$\begin{aligned} \frac{1}{n-1} &= \frac{1/n}{1-1/n} &&= \frac{1}{n}\left(1 + \frac{1}{n} + \frac{1}{n^2} + \frac{1}{n^3} + \ldots\right) \\ \frac{1}{(n-1)^2} &= \frac{1/n^2}{(1-1/n)^2} &&= \frac{1}{n^2}\left(1 + \frac{2}{n} + \frac{3}{n^2} + \frac{4}{n^3} + \ldots\right) \\ \frac{1}{(n-1)^3} &= \frac{1/n^3}{(1-1/n)^3} &&= \frac{1}{n^3}\left(1 + \frac{3}{n} + \frac{6}{n^2} + \frac{10}{n^3} + \ldots\right) \end{aligned}$$

ce qui permet de réécrire $E\left(T_{(.)}\right)$:

$$E\left(T_{(.)}\right) = \theta + \frac{a_1}{n} + \frac{a_1 + a_2}{n^2} + \frac{a_1 + 2a_2 + a_3}{n^3} + \cdots$$

Considérons $\alpha T + \beta T_{(.)}$; nous voyons qu'il est nécessaire de prendre $\alpha + \beta = 1$ pour que l'estimateur $\alpha T + \beta T_{(.)}$ converge vers θ lorsque $n \to \infty$.

Prenons $\alpha + \beta = 1$ et $\alpha = un + v$ (linéaire en n).

$$E\left(\alpha T + \beta T_{(.)}\right) = \theta + \frac{1}{n} a_1 (1-u) + \frac{1}{n^2}\left(-v a_1 + a_2\right) + \cdots$$

Il suffit de prendre $u = 1$ pour éliminer les termes en $\frac{1}{n}$. Si le rapport $\frac{a_1}{a_2}$ était connu $\alpha = n - \frac{a_2}{a_1}$ permettrait d'éliminer aussi les termes en $\frac{1}{n^2}$. Dans le cas du Jackknife, on prend $v = 0$.

1.5 . Variantes sur le Jackknife

a élimination des termes en $1/n^2$

Essayons de construire un estimateur dont l'espérance soit de la forme : $\theta + o\left(n^{-3}\right)$.

Ce problème a été résolu par Gray et Schucany [6]. L'idée est la suivante : pour éliminer les termes en $\frac{1}{n}$ il faut une combinaison de T et $T_{(.)}$, pour éliminer les termes en $\frac{1}{n}$ et $\frac{1}{n^2}$, nous utiliserons une combinaison entre $T, T_{(.)}$ et $T_{(..)}$ où :

$$T_{(..)} = \frac{1}{C_n^2} \sum_{i,j} T_{(-ij)}$$

$T_{(-ij)}$ étant l'estimateur obtenu en éliminant X_i et X_j du n-échantillon $(X_1, X_2, \ldots X_n)$.

Autrement dit :

$$E\left(T_{(-ij)}\right) = E(T..) = \theta + \frac{a_1}{n-2} + \frac{a_2}{(n-2)^2} + \cdots$$

L'estimateur cherché $\tilde{\tilde{T}}$ s'écrit :

$$\tilde{\tilde{T}} = \frac{\begin{vmatrix} T & T_{(.)} & T_{(..)} \\ \frac{1}{n} & \frac{1}{n-1} & \frac{1}{n-2} \\ \frac{1}{n^2} & \frac{1}{(n-1)^2} & \frac{1}{(n-2)^2} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ \frac{1}{n} & \frac{1}{n-1} & \frac{1}{n-2} \\ \frac{1}{n^2} & \frac{1}{(n-1)^2} & \frac{1}{(n-2)^2} \end{vmatrix}}.$$

Après calcul des deux déterminants, nous obtenons :

$$\tilde{\tilde{T}} = \frac{n^2 T - 2(n-1)^2 T_{(.)} + (n-2)^2 T_{(..)}}{2}.$$

Nous vérifions que : $E(\tilde{\tilde{T}}) = \theta + o\left(n^{-3}\right)$.

Remarque :

Nous pourrions évidemment éliminer les termes en $\frac{1}{n^3}, \frac{1}{n^4}, \ldots$ Cette démarche peut paraître alléchante, mais il faut tenir compte :

- des coûts de calcul qui augmentent,
- de l'évolution de la précision des estimateurs.

b le Jackknife de groupe

Admettons que n soit égal au produit de g par h (g et h entiers), nous diviserons le n-échantillon en g sous-ensembles de h-échantillons.

Deux possibilités nous sont offertes pour calculer un estimateur du type Jackknife.

1 : Nous pouvons considérer que les g sous-ensembles déterminent une partition du n-échantillon,

$$\underbrace{x_1\ x_2 \dots x_h}_{g_1} \quad \underbrace{x_{h+1} \dots x_{h+h}}_{g_2} \quad \dots$$

et définir un estimateur du type Jackknife :

$$\breve{T} = gT - (g-1)\breve{T}_{(.)}$$

avec : $\breve{T}_{(.)} = \dfrac{1}{g}\sum_{i=1}^{g} \breve{T}_{(-i)}$

et $\breve{T}_{(-i)}$ l'estimateur construit en éliminant les observations du groupe i

Calcul de $E\left(\breve{T}_{(.)}\right)$:

$$E\left(\breve{T}_{(-i)}\right) = E\left(\breve{T}_{(.)}\right) = \theta + \frac{a_1}{n-h} + \frac{a_2}{(n-h)^2} + \dots$$

$$\begin{aligned}
E\left(\breve{T}_{(.)}\right) &= \theta + \frac{a_1}{h(g-1)} + \frac{a_2}{(h(g-1))^2} + \dots \\
E(T) &= \theta + \frac{a_1}{hg} + \frac{a_2}{(hg)^2} + \dots \\
E(\breve{T}) &= \theta + g\left(\frac{a_2}{(hg)^2} + \frac{a_3}{(hg)^3} + \dots\right) - (g-1)\left(\frac{a_2}{(h(g-1))^2} + \frac{a_3}{(h(g-1))^3} + \dots\right)
\end{aligned}$$

Les termes en a_1 disparaissent :

$$E(\breve{T}) = \theta + a_2\left(\frac{g}{n^2} - \frac{g-1}{(n-1)^2}\right) + a_3\left(\frac{g}{n^3} - \frac{g-1}{(n-h)^3}\right) + \dots$$

Etudions le terme en a_2 :

$$\begin{aligned}
\frac{g}{n^2} - \frac{g-1}{(n-1)^2} &= \frac{g}{n^2} - (g-1)\left(\frac{1}{n^2} + \frac{2h}{n^3} + \frac{3h^2}{n^4} + \dots\right) \\
&= -\frac{1}{n^2} - \frac{h}{n^3} + \dots
\end{aligned}$$

Nous obtenons un estimateur $\breve{T}$ tel que $E(\breve{T}) = \theta + o\left(\frac{1}{n^2}\right)$

Quenouille [13] fut le premier à utiliser une telle méthode dans une étude de séries temporelles en prenant le cas où $g = 2$.

2 : Nous pouvons calculer tous les estimateurs construits sur $(n-h)$ observations

Cette méthode semble plus logique, mais est beaucoup plus coûteuse en temps de calcul.
Ces estimateurs seront notés $\breve{T}_{[-h]}$.
Leur moyenne sera égale à : $\breve{T}_{[.]} = \frac{1}{C_n^h} \sum_h \breve{T}_{[-h]}$
Nous construisons ensuite :

$$\breve{T} = gT - (g-1)\breve{T}_{[.]}$$

et nous pouvons montrer :

$$E(\breve{T}) = \theta + o\left(n^{-2}\right)$$

<u>Conclusion :</u>

A ce niveau (celui du biais), il ne semble pas qu'il y ait intérêt à utiliser des Jackknife de groupe, sauf (peut-être) en ce qui concerne le repérage de sous-ensembles de valeurs aberrantes ou bien si l'on désire faire du "Jackknife" à partir d'un grand échantillon.

Variance du Jackknife

2.1 Variance de l'estimateur du Jackknife

Tukey a introduit la notion de "pseudo-valeurs" :

$$\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$$

l'estimateur du Jackknife apparaissant comme la moyenne de ces "pseudo-valeurs" :

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_{(i)}$$

Il a supputé, semble-t-il par analogie avec la variance d'une moyenne, que :

$$\boxed{s^2 = \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2}$$

était une estimation de la variance de T, et de $\tilde{T}$ l'estimateur du Jackknife.

Par ailleurs, il a fait l'hypothèse que :

$$\frac{\tilde{T} - \theta}{S} \quad \text{où} \quad S^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{T}_{(i)} - \tilde{T}\right)^2$$

suivait une loi de Student à (n-1) ddl.

Tukey part de l'idée que les "pseudo-valeurs" $\tilde{T}_{(i)}$ (variables aléatoires qui prennent pour valeur $\tilde{\theta}_{(i)}$) sont indépendantes, ou faiblement corrélées.

Démonstration :

Si nous calculons les covariances entre "pseudo-valeurs", nous obtenons :

$$\begin{aligned}\mathrm{Cov}\left(\tilde{T}_{(i)}, \tilde{T}_{(i')}\right) &= \mathrm{Cov}\left(nT - (n-1)T_{(-i)}, nT - (n-1)T_{(-i')}\right)\\ &= n^2\mathrm{Var}(T) - 2n(n-1)\mathrm{Cov}\left(T, T_{(-i)}\right) + (n-1)^2\mathrm{Cov}\left(T_{(-i)}, T_{(-i')}\right)\end{aligned}$$

Nous admettrons que les observations sont indépendantes et que la statistique est symétrique : $\mathrm{Cov}\left(T, T_{(-i)}\right) = \mathrm{Cov}\left(T, T_{(-i')}\right)$.

Considérons une statistique de la forme :

$$T(X_1, X_2, \ldots X_n) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Si nous posons $\gamma = \mathrm{Var}(f(X))$, nous avons :

$$\mathrm{Var}(T) = \frac{1}{n}\gamma \quad \mathrm{Cov}\left(T, T_{(-i)}\right) = \frac{1}{n}\gamma \quad \mathrm{Cov}\left(T_{(-i)}, T_{(-i')}\right) = \frac{n-2}{(n-1)^2}\gamma$$

et nous trouvons :

$$\mathrm{Cov}\left(\tilde{T}_{(i)}, \tilde{T}_{(i')}\right) = 0$$

Nous obtenons aussi :

$$\mathrm{Var}\left(\tilde{T}_{(i)}\right) = \gamma$$

En effet :

$$\begin{aligned}
\tilde{T}_{(i)} &= nT - (n-1)T_{(-i)} \\
\mathrm{Var}\left(\tilde{T}_{(i)}\right) &= n^2\mathrm{Var}(T) + (n-1)^2\mathrm{Var}\left(T_{(-i)}\right) - 2n(n-1)\mathrm{Cov}\left(T, T_{(-i)}\right) \\
&= n^2\frac{\gamma}{n} + (n-1)^2\frac{\gamma}{n-1} - 2n(n-1)\frac{\gamma}{n} \\
&= \gamma
\end{aligned}$$

Nous pouvons donc prendre $\frac{1}{n}\sum\left(\tilde{T}_{(i)} - \tilde{T}\right)^2$ comme estimateur de γ

et $\frac{1}{n(n-1)}\sum\left(\tilde{T}_{(i)} - \tilde{T}\right)^2$ comme estimateur de $\mathrm{Var}\left(\tilde{\mathrm{T}}\right)$

<u>Conclusion :</u>

Il est évident que les statistiques de ce type sont peu nombreuses $\left(\bar{X} = \frac{1}{n}\sum X_i\right)$, par contre, de nombreuses statistiques peuvent être approchées par des estimateurs de cette forme pour n suffisamment grand.

Remarque 1 :

$$s^2 = \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 = \frac{(n-1)}{n}\sum\left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2$$

En effet :

$$\begin{aligned}
\tilde{\theta}_{(i)} &= n - (n-1)\hat{\theta}_{(-i)} \\
\tilde{\theta} &= n\hat{\theta} - \frac{n-1}{n}\sum\hat{\theta}_{(-i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(.)}
\end{aligned}$$

D'où :

$$\tilde{\theta}_{(i)} - \tilde{\theta} = (n-1)\left(\hat{\theta}_{(.)} - \hat{\theta}_{(-i)}\right)$$

Remarque 2 :

Un autre estimateur de la variance a été proposé :

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \hat{\theta}\right)^2$$

Etudions le lien entre s'^2 et s^2 :

$$\begin{aligned}
s'^2 &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \hat{\theta}\right)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\tilde{\theta}_{(i)} - \tilde{\theta} + \tilde{\theta} - \hat{\theta}\right)^2 \\
&= \frac{1}{n(n-1)}\sum\left\{\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 + \left(\tilde{\theta} - \hat{\theta}\right)^2 + 2\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)\left(\tilde{\theta} - \hat{\theta}\right)\right\} \\
&= \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 + \frac{n}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)^2 + \frac{2}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right) \\
&= \frac{1}{n(n-1)}\sum\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 + \frac{n}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)^2 + \frac{2n}{n(n-1)}\left(\tilde{\theta} - \hat{\theta}\right)\left(\tilde{\theta} - \tilde{\theta}\right)
\end{aligned}$$

D'où :

$$\boxed{s'^2 = s^2 + \frac{1}{n-1}\left(\tilde{\theta} - \hat{\theta}\right)^2}$$

Cet estimateur est considéré comme plus "conservateur" car le terme $\left(\tilde{\theta} - \hat{\theta}\right)^2$ est toujours positif, et donc :

$$s'^2 \geq s^2$$

Remarque 3 :

L'hypothèse avancée par Tukey permet de définir un intervalle de confiance, au niveau $(1-\alpha)$ du paramètre θ :

$$\theta \in \left[\tilde{\theta} - t_{\alpha/2;(n-1)}s \ ; \ \tilde{\theta} + t_{\alpha/2;(n-1)}s\right]$$

Remarque 4 :

La conjecture de Tukey a été infirmée par des contre-exemples, dont certains seront présentés dans les paragraphes suivants.

Bien qu'il y ait une certaine robustesse de la loi de Student vis-à-vis de la non-indépendance des échantillons, il faut faire attention au cas de "pseudo-valeurs" discrètes, en particulier dans le cas d'estimation de la médiane ou d'autres statistiques d'ordre.

Une règle pour réduire les degrès de liberté de T s'avère nécessaire : celle proposée par Mosteller et Tukey [12] est simple et utile :

- compter le nombre de "pseudo-valeurs" réellement différentes ;
- en soustraire un ;
- utiliser le résultat comme degrés de liberté.

2.2 Démarche suivie par Efron

Nous avons vu :

$$s^2 = \frac{1}{n(n-1)} \sum \left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 = \frac{(n-1)}{n} \sum \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2$$

C'est sous la deuxième forme que l'estimation de la variance est utilisée par Efron [4], il décompose la construction de s^2 en deux étapes :

1. $\sum \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2$ qui est construit en n'utilisant que des $(n-1)$ -échantillons, est considéré comme une estimation de γ_{n-1} la variance de T_{n-1} .
2. $\dfrac{n-1}{n}$ est alors considéré comme un coefficient correcteur qui permet de passer à l'estimation de γ_n.

Efron démontre, par ailleurs :

$$E\left(\sum_{i=1}^{n} \left(T_{(-i)} - T_{(.)}\right)^2\right) \geq \gamma_{n-1}$$

La démonstration est développée en annexe.

Remarque 1 :
Efron ne s'intéresse qu'aux statistiques qui sont des fonctions symétriques des X_i.

Remarque 2 :
Si T est de la forme : $\dfrac{1}{n}\sum_{i=1}^{n} f\left(X_i\right)$, les termes en $\sigma_B^2, \ldots$ (Cf. Annexe 1) sont nuls et $\sum_{i=1}^{n} \left(T_{(i)} - T_{(.)}\right)^2$ est un estimateur sans biais de γ_{n-1}.

Remarque 3 :

Généralement, nous n'avons pas :

$$E\left(\frac{n-1}{n}\sum_{i=1}^{n}\left(T_{(i)}-T_{(.)}\right)^2\right)\geq\gamma_n$$

Cependant cette inégalité est vraie pour les U-statistiques.

Une U-statistique est une statistique de la forme :

$$U_n\left(X_1,X_2,\ldots X_n\right)=\frac{1}{C_n^m}\sum_C g\left(X_{i1},X_{i2},\ldots X_{im}\right)$$

où C indique que la sommation s'effectue sur l'ensemble des combinaisons de m entiers choisis parmi $(1,2,\ldots n)$, de plus g est symétrique et $E(g), E\left(g^2\right),\ldots$ existent.

L'estimateur de la variance, par exemple, est une U-statistique :

$$\begin{aligned}\frac{1}{n-1}\sum\left(X_i-\bar{X}\right)^2 &= \frac{1}{2(n-1)n}\sum_{i,i'}\left(X_i-X_{i'}\right)^2\\ &= \frac{1}{C_n^2}\sum_{i,i'}4\left(X_i-X_{i'}\right)^2\end{aligned}$$

Hoeffding [7] a montré que pour une U-statistique avec $m\leq n-1$, nous avons :

$$\frac{n-1}{n}\text{var}\left(T_{n-1}\right)\geq\text{var}\left(T_n\right)$$

ce qui entraîne :

$$E\left(\frac{n-1}{n}\sum_i\left(T_{(-i)}-T_{(.)}\right)^2\right)\geq\gamma_n$$

2.3 Exemples de variances d'estimateurs

a l'estimateur de la moyenne

Soit $\hat{\theta}$, la valeur prise par l'estimateur de la moyenne. Nous savons que :

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$$

Nous pouvons montrer aisément que les "pseudo-valeurs" et la valeur prise par l'estimateur du Jackknife s'écrivent :

$$\begin{aligned}\tilde{\theta}_{(i)} &= x_i\\ \tilde{\theta} &= \frac{1}{n}\sum\hat{\theta}_{(-i)}=\hat{\theta}\end{aligned}$$

D'où la variance de l'estimateur de la moyenne :

$$\begin{aligned} s^2 &= \frac{1}{n(n-1)} \sum \left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2 \\ &= \frac{1}{n(n-1)} \sum (x_i - \bar{x})^2 \end{aligned}$$

Résultat fort connu !

b l'estimateur fonction de la moyenne

Nous étudierons un estimateur fonction de la moyenne : $T = g\left(\bar{X}\right)$.

Notations : $\mu = E\left(\bar{X}\right) \quad \gamma = var\left(\bar{X}\right)$

Nous admettrons que g est dérivable et $E(g(X)) \simeq g(E(X))$

Nous pouvons obtenir une estimation de $\mathrm{Var}(T)$ par la méthode suivante :

$$\begin{aligned} g(x) &= g(\mu) + (x-\mu)g'(\mu) + o\left((x-\mu)^2\right) \\ E\left(\left(g\left(\bar{X}\right) - g(\mu)\right)^2\right) &\simeq E\left(\left(\bar{X} - \mu\right)^2\right)(g'(\mu))^2 \\ \mathrm{Var}\left(g\left(\bar{X}\right)\right) &= \mathrm{Var}(T) = (g'(\mu))^2 \mathrm{Var}\left(\bar{X}\right) \end{aligned}$$

Remarque :

C'est cette expression que l'on utilise pour stabiliser la variance.
Les transformations : Log, Arcsin$\sqrt{\ }$... étant obtenues en résolvant l'équation différentielle :
$K = (g'(\mu))^2 \mathrm{Var}\left(\bar{X}\right)$ avec $\mathrm{Var}\left(\bar{X}\right) = k\mu^2$; $\mathrm{Var}\left(\bar{X}\right) = p(p-1)\ldots$
K constante positive

Si nous utilisons les "pseudo-valeurs" :

$$\begin{aligned} \hat{\theta}_{(-i)} &= g\left(\frac{n\bar{x} - x_i}{n-1}\right) = g\left(\bar{x} + \frac{\bar{x} - x_i}{n-1}\right) \\ &= g(\bar{x}) + g'(\bar{x})\frac{\bar{x} - x_i}{n-1} + o\left(\left(\frac{\bar{x} - x_i}{n-1}\right)^2\right) \\ \hat{\theta}_{(.)} &= g(\bar{x}) + \sum_i o\left(\left(\frac{\bar{x} - x_i}{n-1}\right)^2\right) \end{aligned}$$

D'où ;

$$
\begin{aligned}
\hat{\theta}_{(-i)} - \hat{\theta}_{(.)} &\simeq g'(\bar{x}) \frac{\bar{x} - x_i}{n-1} \\
\frac{n-1}{n} \sum_i \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)}\right)^2 &\simeq \frac{n-1}{n} (g'(\bar{x}))^2 \sum_i \frac{(\bar{x} - x_i)^2}{(n-1)^2} \\
&= (g'(\bar{x}))^2 \sum_i \frac{(\bar{x} - x_i)^2}{n(n-1)}
\end{aligned}
$$

Nous voyons l'analogie avec l'estimateur obtenu par la première méthode.

c l'estimateur de la variance

Considérons l'estimateur de la variance :

$$V = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Cet estimateur est non biaisé, et l'estimateur du Jackknife $\tilde{V}$ est égal à V.

La valeur prise par cet estimateur s'écrit :

$$\hat{\theta} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Les "pseudo-valeurs" s'écrivent :

$$\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$$

La valeur prise par la variance de l'estimateur est :

$$s^2 = \frac{1}{n(n-1)} \sum \left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)^2$$

Si nous définissons le moment centré d'ordre k :

$$
\begin{aligned}
\mu_k &= E(X - E(X))^k \\
\text{et} \quad \hat{\mu}_k &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k
\end{aligned}
$$

Nous obtenons :

$$s^2 = \frac{n^2}{(n-1)(n-2)^2} \left(\hat{\mu}_4 - \hat{\mu}_2^2\right)$$

Or la vraie variance s'écrit :

$$\mathrm{Var}(V) = \frac{\mu_4 - \frac{n-3}{n-1}\mu_2^2}{n}$$

Soit A le coefficient d'aplatissement de Pearson, nous avons : $A = \frac{\mu_4}{\mu_2^2}$.

Nous pouvons réécrire les valeurs des variances :

$$\begin{aligned} s^2 &= \frac{n^2}{(n-1)(n-2)^2}(A-1)\hat{\mu}_2^2 \\ \mathrm{Var}(V) &= \frac{A - \frac{n-3}{n-1}}{n}\mu_2^2 \end{aligned}$$

Nous voyons ainsi que le rapport $\frac{s^2}{\mathrm{Var}(V)}$ dépend de la taille de l'échantillon et du coefficient d'aplatissement :

Le tableau suivant présente quelques rapports, pour différentes valeurs de n et A :

n \ A	1,5	1,8	3
5	1,74	2,14	2,78
10	1,20	1,36	1,56
15	1,11	1,21	1,33
20	1,07	1,15	1,23
100	1,01	1,03	1,04

Nous voyons que ce rapport diminue et qu'il est inversement proportionnel à n et A.

Rappel :

Pour une loi Normale $\mathcal{N}(0,1)$, nous avons : $\mu_2 = 1$, $\mu_4 = 3$, $A = 3$.
Pour une loi Uniforme comprise entre $[a-h, a+h]$: $\mu_k = \frac{1}{k+1}h^k$, $A = \frac{9}{5}$.

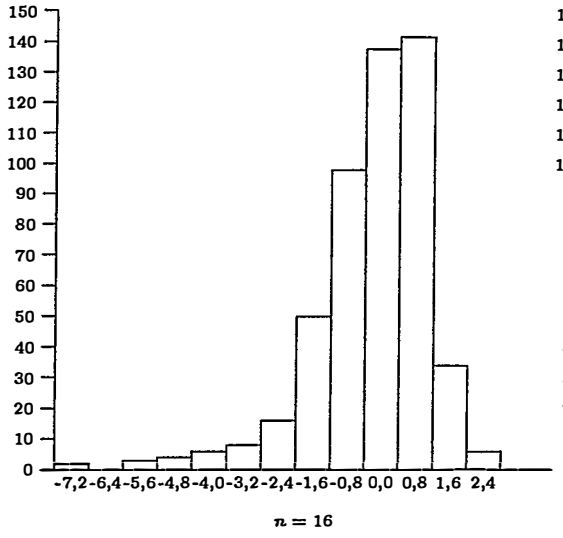
d l'estimateur de l'écart-type

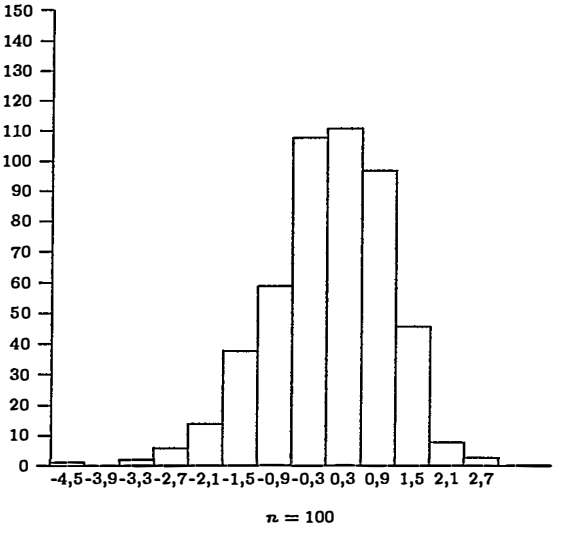
Nous reprenons les simulations effectuées au premier chapitre et, pour chacune d'entre-elles, nous calculons :

$$t = \frac{\tilde{s} - \sqrt{10}}{\sqrt{\frac{1}{n(n-1)}\sum\left(\tilde{s}_{(i)} - \tilde{s}\right)^2}}$$

Nous obtenons ainsi 500 valeurs de t : $t_1, t_2, \ldots, t_{500}$ dont nous pouvons étudier la distribution.

En observant les histogrammes, nous constatons pour $n = 16$, 50 et 100 que les hypothèses formulées par Tuckey semblent réalistes (évolution vers une loi Normale).





e l'estimateur d'un quotient

Reprenons l'exemple présenté au chapitre précédent, concernant l'estimation de la surface boisée nationale.

La variance de l'estimateur du Jackknife est égale à :

$$\begin{aligned} s^2 &= \frac{1}{n(n-1)} \sum \left(\tilde{q}_{(i)} - \tilde{q}\right)^2 = \frac{1}{10 * 9} 1150,128 \\ &= 12,779 \end{aligned}$$

Pour valider ce résultat, nous avons tiré aléatoirement 100 échantillons de 10 départements parmi les 90. Pour chacun de ces échantillons, nous avons calculé l'estimateur du quotient et celui du Jackknife.

Les résultats sont présentés dans le tableau suivant :

Estimateur	Moyenne	Minimum	Maximum	Variance
Q	24,754	13,86	39,78	16,70
$\tilde{Q}$	24,775	13,80	40,14	17,33

L'estimateur du Jackknife est légèrement plus proche de la vraie valeur que l'estimateur du quotient.
Les variances observées sont semblables à l'estimation de la variance proposée par Tukey.

f méthode de "capture-recapture"

Soit N le nombre d'individus dans une population; ce nombre est inconnu, voire inobservable.
Nous supposons qu'il existe deux listes ou structures, qui représentent chacune une partie de la population. Ces deux listes sont indépendantes.

		Seconde Liste Présent	Seconde Liste Absent	
Première Liste	Présent	N_{11}	N_{12}	$N_{1.}$
	Absent	N_{21}	—	
		$N_{.1}$		

Le nombre N_{22} est inobservable, et par conséquent la taille de la population N aussi.

Si nous supposons que ces données proviennent d'une loi multinomiale, la valeur prise par l'estimateur du maximum de vraisemblance s'écrit :

$$\hat{n} = \frac{n_{1.}n_{.1}}{n_{11}}$$

L'exemple présenté concerne l'estimation de la taille de la population agée de 14 à 64 ans des Etats-Unis [16].

La première liste représente une enquête de la population en février 1978, la seconde liste représente les impôts perçus durant l'année 1978. Ces listes seront notées respectivement "E" et "I".

Le plan d'échantillonnage de l'enquête est complexe. Les résultats proviennent de huit sous-échantillons, qui fournissent chacun des estimations de la population nationale. Les données provenant de "I" sont connues avec exactitude.

Ainsi, nous observons :

		I Présent	I Absent	
E	Présent	$\hat{n}_{11i}$	$\hat{n}_{12i}$	$\hat{n}_{1.i}$
	Absent	$\hat{n}_{21i}$	—	
		$n_{.1}$		

pour $i = 1, 2, \ldots 8$ qui correspondent aux groupes sur lesquels ont été faites les estimations. $n_{.1}$ n'est pas une estimation, mais un dénombrement.

Nous obtenons ainsi, le tableau global :

		I		
		Présent	Absent	
E	Présent	$\hat{n}_{11}$	$\hat{n}_{12}$	$\hat{n}_{1.}$
	Absent	$\hat{n}_{21}$	–	
		$n_{.1}$		

avec :

$$\hat{n}_{11} = \frac{1}{8}\sum_{i=1}^{8}\hat{n}_{11i}, \quad \hat{n}_{12} = \frac{1}{8}\sum_{i=1}^{8}\hat{n}_{12i},$$
$$\hat{n}_{21} = \frac{1}{8}\sum_{i=1}^{8}\hat{n}_{21i}, \quad \hat{n}_{1.} = \frac{1}{8}\sum_{i=1}^{8}\hat{n}_{1.i}.$$

Les données sont présentées dans le tableau suivant :

Groupe	$\hat{n}_{11h}$	$\hat{n}_{1.h}$
1	107.285.040	133.399.520
2	105.178.160	132.553.952
3	110.718.448	139.055.744
4	103.991.496	132.390.240
5	106.818.488	131.627.520
6	106.636.928	133.095.536
7	105.338.552	133.324.528
8	103.349.328	131.061.688
Moyenne	106.164.555	133.313.591

Le nombre d'individus "présents" dans la liste "I" est : $n_{.1} = 115.090.300$

L'estimateur de la méthode de "capture-recapture" est :

$$\hat{n} = \frac{\hat{n}_{1.}n_{.1}}{\hat{n}_{11}} = \frac{(133.313.591)(115.090.300)}{106.164.555} = 144.521.881$$

Nous éliminons le $i^{ème}$ groupe, et nous calculons les nouvelles estimations :

$$\hat{n}_{(-i)} = \frac{\hat{n}_{1.(-i)}n_{.1}}{\hat{n}_{11(-i)}}$$

avec:

$$\hat{n}_{11(-i)} = \frac{1}{7}\sum_{i'\neq i}\hat{n}_{11i'}$$
$$\hat{n}_{1.(-i)} = \frac{1}{7}\sum_{i'\neq i}\hat{n}_{1.i'}$$

Les "pseudo-valeurs" s'écrivent :

$$\tilde{n}_{(i)} = 8\hat{n} - 7\hat{n}_{(-i)}$$

Les résultats sont présentés dans le tableau suivant :

Groupe	$\hat{n}_{(-i)}$	$\tilde{n}_{(i)}$
1	144.726.785	143.087.553
2	144.447.797	145.040.467
3	144.518.186	144.547.744
4	144.243.095	146.473.380
5	144.910.512	141.801.461
6	144.647.594	143.641.892
7	144.359.733	145.656.914
8	144.323.897	145.907.770
Moyenne		144.519.648

L'estimateur du Jackknife est égal à :

$$\tilde{n} = \frac{1}{8}\sum_{i=1}^{8}\tilde{n}_{(i)} = 144.519.648$$

Nous pouvons calculer la variance et l'écart-type de l'estimateur du Jackknife :

$$\begin{aligned} s^2 &= \frac{1}{8*7}\sum_{i=1}^{8}\left(\tilde{n}_{(i)} - \tilde{n}\right)^2 = 3,1284*10^{11} \\ s &= 559.321 \end{aligned}$$

ainsi que la variance "conservatrice" :

$$s'^2 = \frac{1}{8*7}\sum_{i=1}^{8}\left(\tilde{n}_{(i)} - \hat{n}\right)^2 = 3,1284*10^{11}$$

2.4 Contre-exemple des spéculations de Tuckey

Reprenons l'exemple de l'estimateur de la borne d'une loi uniforme présentée au paragraphe 1.3.c.

Soit une variable aléatoire X distribuée de façon uniforme sur $[0, \theta]$ et le n -échantillon $(X_1, X_2, \ldots X_n)$ issu de la variable parente X.

Nous étudierons la distribution de :

$$t = \frac{\tilde{\theta} - \theta}{\sqrt{\mathrm{Var}\left(\tilde{\theta}\right)}}$$

Nous utiliserons les notations habituelles, $X_{(i)}$ désignant la variable X_j de rang i :

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(i)} \leq \cdots \leq X_{(n)}$$

L'estimateur du Jackknife s'écrit :

$$\tilde{\theta} = X_{(n)} + \frac{n-1}{n}\left(X_{(n)} - X_{(n-1)}\right)$$

avec :

$$\mathrm{Var}\left(\tilde{\theta}\right) = \frac{1}{n(n-1)} \sum \left(\tilde{\theta}_i - \tilde{\theta}\right)^2$$

Nous poserons $\theta = 1$, et nous étudierons la variable T qui s'écrit :

$$T = \sqrt{\frac{n-1}{n}} Z + \sqrt{\frac{n-1}{n}} \quad \text{avec} \quad Z = \frac{X_{(n)} - 1}{X_{(n)} - X_{(n-1)}}$$

Etudions la loi du couple $\left(X_{(n)}, X_{(n-1)}\right)$:

$$F(x, y) = P\left(X_{(n-1)} < x \text{ et } X_{(n)} < y\right)$$

Si $x > y$, nous avons $F(x, y) = 1$

Si $x < y$:

En utilisant la relation : $P(B) = P(B \cap A) + P\left(B \cap \bar{A}\right)$
et en posant : $A = \left[X_{(n)} < y\right] \quad \bar{A} = \left[X_{(n)} > y\right] \quad \text{et} \quad B = \left[X_{(n-1)} < x\right]$

Nous obtenons :

$$\begin{aligned}
F(x,y) &= P\left(X_{(n-1)} < x \text{ et } X_{(n)} < y\right) \\
&= P\left(X_{(n-1)} < x\right) + P\left(X_{(n-1)} < x \text{ et } X_{(n)} > y\right) \\
P\left(X_{(n-1)} < x\right) &= nx^{n-1}(1-x) + x^n \\
P\left(X_{(n-1)} < x \text{ et } X_{(n)} > y\right) &= nx^{n-1}y \\
\text{soit} \qquad F(x,y) &= x^n(1-n) + nx^{n-1}y \qquad \text{si } x < y
\end{aligned}$$

En conclusion, si $f(x,y)$ désigne la densité du couple (X,Y) :

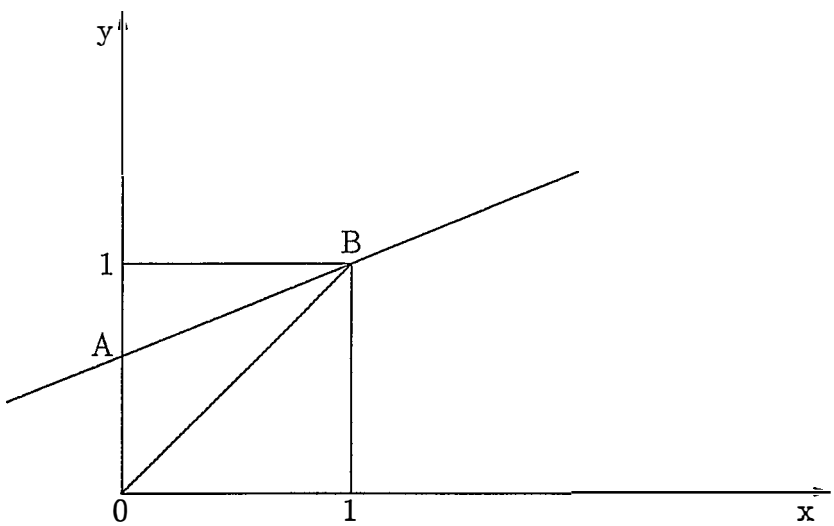
$$f(x,y) = \frac{\partial F(x,y)}{\partial x \partial y} = \begin{cases} (n-1)nx^{n-2} & \text{si } x < y \\ 0 & \text{si } x > y \end{cases}$$

Etudions la loi de $Z = \dfrac{X_{(n)} - 1}{X_{(n)} - X_{(n-1)}}$

Nous avons :

$$F(z) = P(Z < z) = \begin{cases} 1 & \text{si } z < 0 \quad \text{car } X_{(n)} < 1 \\ P\left(X_{(n)}(1-z) + X_{(n-1)}z < 1\right) & \text{si } z > 0 \end{cases}$$

Etude des droites $y(1-z) + zx = 1$ (z étant un paramètre négatif) : ces droites forment un faisceau passant par le point $(1,1)$ et $P(Z < z)$ est la probabilité que (X,Y) appartiennent au triangle $(O\ A\ B)$ (voir figure) :



$$P(Z < z) = \int\int_{(OAB)} (n-1)nx^{n-2}\mathrm{d}x\mathrm{d}y = \frac{1}{1-z}$$

<u>Conclusion :</u>

$$F(z) = \begin{cases} 1 & \text{si } z > 0 \\ \dfrac{1}{1-z} & \text{si } z < 0 \end{cases} \qquad f(z) = \begin{cases} 0 & \text{si } z > 0 \\ \dfrac{1}{(1-z)^2} & \text{si } z < 0 \end{cases}$$

Etudions la loi de T :

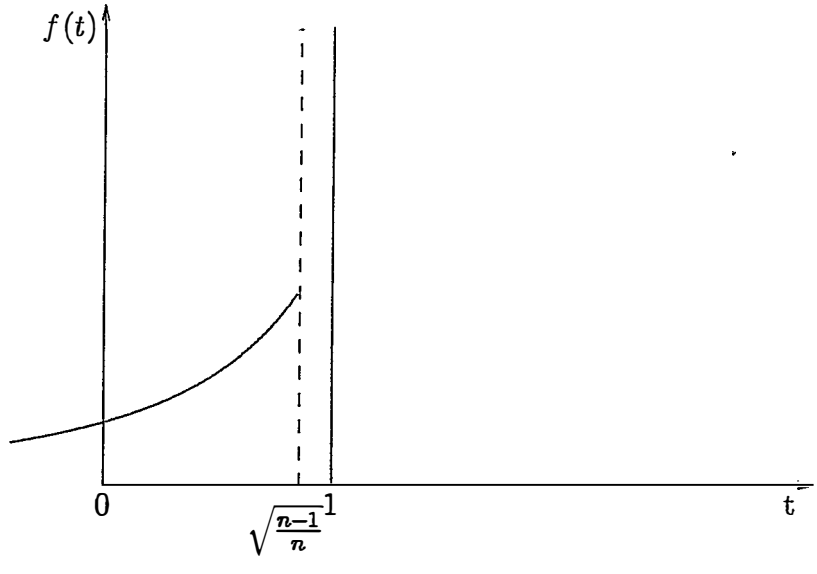
$$T = \sqrt{\frac{n}{n-1}}Z + \sqrt{\frac{n-1}{n}} \quad \text{avec} \quad Z = \frac{X_{(n)} - 1}{X_{(n)} - X_{(n-1)}}$$

Nous déduisons aisément la loi de T de la loi de Z :

$$\begin{aligned} F(t) &= P(T < t) = P\left(Z < \sqrt{\tfrac{n-1}{n}}t - \tfrac{n-1}{n}\right) \\ &= \begin{cases} 1 & \text{si } t > \sqrt{\frac{n-1}{n}} \simeq 1 \\ \dfrac{n}{(2n-1) - t\sqrt{n(n-1)}} & \text{si } t < \sqrt{\frac{n-1}{n}} \end{cases} \end{aligned}$$

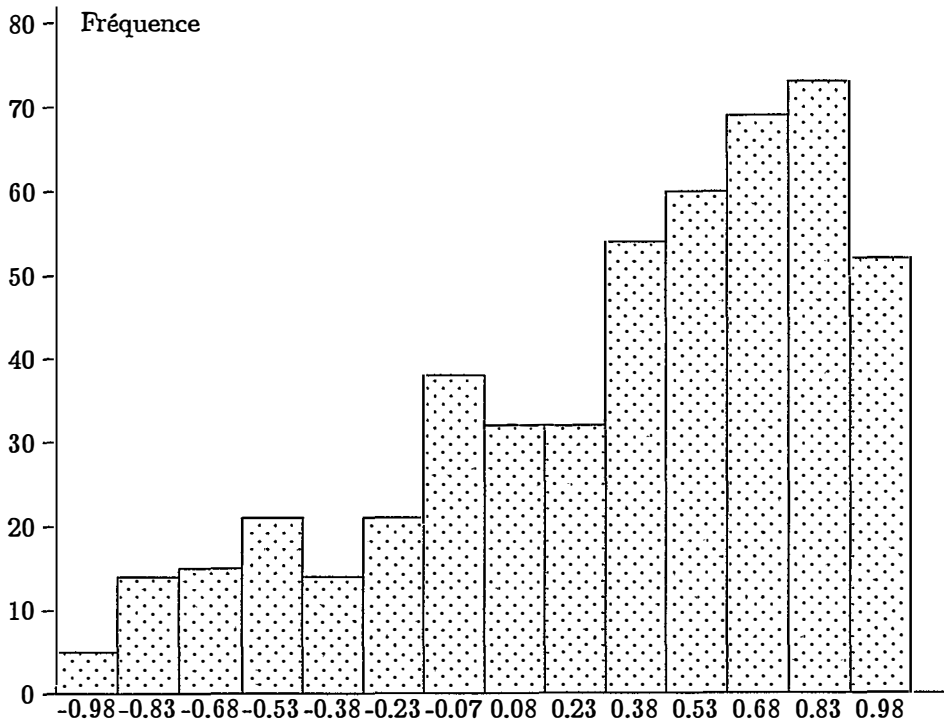
$$f(t) = \begin{cases} 0 & \text{si } t > \sqrt{\frac{n-1}{n}} \\ \dfrac{n\sqrt{n(n-1)}}{\left((2n-1) - t\sqrt{n(n-1)}\right)^2} & \text{si } t < \sqrt{\frac{n-1}{n}} \end{cases}$$

Le graphe suivant représente l'allure de la fonction f(t) :



La distribution de T est loin d'une distribution du type de loi normale.

Ce résultat est conforté par la simulation de 500 échantillons de taille $n = 50$, dont nous avons représenté la distribution entre -1 et +1, sur l'histogramme suivant :


Fréquence
80
70
60
50
40
30
20
10
0
-0.98 -0.83 -0.68 -0.53 -0.38 -0.23 -0.07 0.08 0.23 0.38 0.53 0.68 0.83 0.98

Le Bootstrap

3.1 Introduction

Avant d'aborder la présentation du Bootstrap nous allons évoquer deux procédures communément utilisées par les biométriciens.

a Substitution de la distribution empirique à la distribution théorique

Soit X une variable aléatoire de loi inconnue $\mathcal{F}$ et l'échantillon $(x_1, x_2, \ldots x_n)$ correspondant à la réalisation de n variables aléatoires $(X_1, X_2, \ldots X_n)$ indépendantes et de loi $\mathcal{F}$.

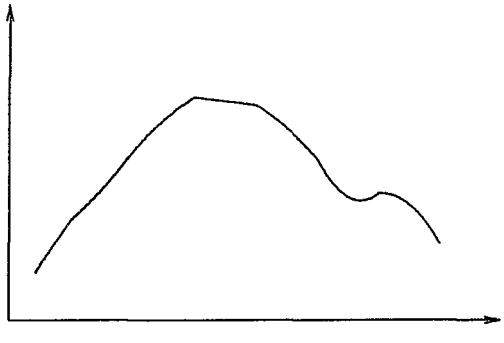
Dans la suite, $\mathcal{F}$ désignera la fonction de répartition de X, soit :

$$\mathcal{F}(t) = P(X \leq t)$$

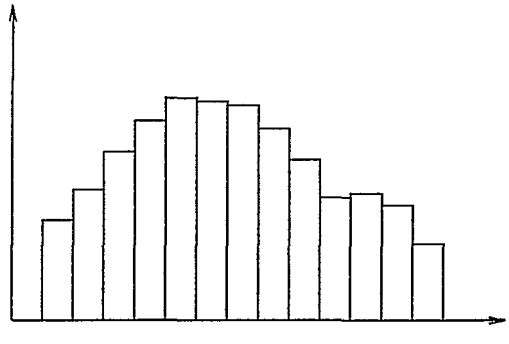
Classiquement, nous approchons $\mathcal{F}$ par la construction d'un histogramme sur les x_i. Nous substituons à la variable X la variable X^* de loi $\mathcal{F}^*$ avec :

$$P\left(c_i < X^* \leq c_{i+1}\right) = \frac{n_i}{n}$$

c_i et c_{i+1} étant les limites de la $i^{ème}$ classe de l'histogramme et n_i le nombre de x_j compris entre c_i et c_{i+1}.



densité de $\mathcal{F}$



histogramme de x_i

Si n est grand, nous pouvons améliorer l'histogramme en opérant par lissage ; par contre si n est faible, nous substituons à X la variable X^* telle que, en supposant les x_i différents :

$$P\left(X^* = x_i\right) = \frac{1}{n}$$

$$\boxed{X \text{ de loi inconnue } \mathcal{F} \;\rightarrow\; X^* \text{ de loi empirique } \mathcal{F}^*}$$

b Utilisation de la simulation

Soit X de loi connue $\mathcal{F}$ et un échantillon de n variables aléatoires $(X_1, X_2, \ldots X_n)$ indépendantes et de loi $\mathcal{F}$. Nous considérons la statistique :

$$T_n = T\left(X_1 . X_2, \ldots X_n\right)$$

Nous nous intéressons à la loi de T_n et aux valeurs des principaux paramètres associés $(E\left(T_n\right)\ ;\ Var\left(T_n\right)\ ;\ \cdots)$.

Deux approches sont envisageables :

le calcul direct

A titre d'exemple, nous pouvons calculer :

$$\begin{aligned}
P\left(T_n < t\right) &= \int \cdots \int_{[T(x_1, x_2, \ldots x_n) < t]} d\mathcal{F}\left(x_1\right) d\mathcal{F}\left(x_2\right) \cdots d\mathcal{F}\left(x_n\right) \\
E\left(T_n\right) &= \int \cdots \int T\left(x_1, x_2, \ldots x_n\right) d\mathcal{F}\left(x_1\right) d\mathcal{F}\left(x_2\right) \cdots d\mathcal{F}\left(x_n\right) \\
E\left(T_n^2\right) &= \int \cdots \int \left[T\left(x_1, x_2, \ldots x_n\right)\right]^2 d\mathcal{F}\left(x_1\right) d\mathcal{F}\left(x_2\right) \cdots d\mathcal{F}\left(x_n\right)
\end{aligned}$$

la simulation

Puisque $\mathcal{F}$ est connue, nous pouvons construire par simulation B répétitions de $(X_1, X_2, \ldots X_n)$, et calculer pour chacune d'entre elles les valeurs correspondantes de T_n :

$$1 \leq j \leq B \qquad \left(x_{1j}, x_{2j}, \ldots x_{nj}\right) \qquad t_j = T\left(x_{1j}, x_{2j}, \ldots x_{nj}\right)$$

A partir de $(t_1, t_2, \ldots, t_j, \ldots t_B)$ on peut alors estimer la distribution de T_n et les valeurs des principaux paramètres associés, par exemple :

$$E\left(T_n\right) = \frac{1}{N} \sum_{j=1}^{B} t_j = \bar{t} \qquad \mathrm{Var}\left(T_n\right) = \frac{1}{B-1} \sum_{j=1}^{B} \left(t_j - \bar{t}\right)^2 \qquad \text{etc } \cdots$$

c La méthode du Bootstrap

Soit X de loi inconnue $\mathcal{F}$ et un échantillon $(x_1, x_2, \ldots x_n)$ se présentant comme la réalisation de n variables aléatoires indépendantes $(X_1, X_2, \ldots X_n)$ avec X_i de loi $\mathcal{F}$. Nous considérons la statistique T_n :

$$T_n = T(X_1, X_2, \ldots X_n)$$

qui est associée à l'estimation d'un paramètre θ :

$$\hat{\theta} = T(x_1, x_2, \ldots x_n)$$

Le Bootstrap consiste à remplacer, comme nous l'avons illustré au paragraphe a, X par X^* et la statistique T_n par T_n^* :

$$T_n^* = T(X_1^*, X_2^*, \ldots X_n^*)$$

A l'identique de la procédure décrite au paragraphe b, on peut ensuite calculer soit directement, soit par simulation, la loi de T_n^* et des paramètres associés. En particulier $E(T_n^*)$ est pris comme estimation de θ :

$$\hat{\theta}_B = E(T_n^*)$$

On parle d'estimateur du Bootstrap.

La précision de cet estimateur est mesurée par $\mathrm{Var}(T_n^*)$:

$$\mathrm{Var}(T_n^*) = E\left(T_n^{*2}\right) - \left[E(T_n^*)\right]^2$$

$\mathrm{Var}(T_n^*)$ est aussi utilisée comme estimation de la variance de T_n, mais on peut lui préférer :

$$E\left(T_n^{*2}\right) - \hat{\theta}^2$$

3.2 Exemples

On va proposer deux exemples : le premier qui concerne la médiane permettra d'illustrer une approche directe de la loi de T_n, le second relatif à la variance de l'estimateur du coefficient de corrélation fera appel à la simulation.

a La médiane

Soit X une variable aléatoire de loi inconnue $\mathcal{F}$. La médiane est le paramètre θ tel que :

$$P(X \leq \theta) = 0,5$$

On considère $(x_1, x_2, \ldots x_i, \ldots x_n)$, une réalisation de n variables aléatoires

$(X_1, X_2, \ldots X_i, \ldots X_n)$ indépendantes et de même loi $\mathcal{F}$.

On note $\left(x_{(1)}, x_{(2)}, \ldots x_{(i)}, \ldots x_{(n)}\right)$ le n échantillon ordonné et l'on s'intéresse à la statistique : $T_n = T\left(X_1, X_2, \ldots X_n\right)$ estimateur de la médiane qui prend pour valeur $\hat{\theta}$ avec :

$$\hat{\theta} = T\left(x_1, x_2, \ldots x_n\right) = \begin{cases} x_{(m+1)} & \text{si } n = 2m+1 \\ \dfrac{x_{(m)} + x_{(m+1)}}{2} & \text{si } n = 2m \end{cases}$$

Comme nous l'avons décrit au paragraphe précédent, nous remplaçons X par X^* avec :

$$P\left(X^* = x_i\right) = \frac{n_i}{n} \qquad (n_i \text{ nombre d'occurrences de } x_i)$$

et :

$$T_n \text{ par } T_n^* = T_n\left(X_1^*, X_2^*, \ldots X_n^*\right)$$

On s'intéresse alors à la loi de T_n^* qui se substitue à la loi inconnue de T_n.

Cas général : Loi de T_n^* lorsque $n = 2m+1$

Rappel : on possède une réalisation $(x_1, x_2, \ldots x_n)$ de $(X_1, X_2, \ldots X_n)$ et l'on note X^* la variable aléatoire telle que :

$$P\left(X^* = x_j\right) = \frac{1}{n} \qquad \text{(on suppose les } x_j \text{ tous différents)}$$

Notons $(x_1^*, x_2^*, \ldots x_n^*)$ une réalisation de $(X_1^*, X_2^*, \ldots X_n^*)$ on a :

$$\forall i \quad \exists j \quad \text{tel que } x_i^* = x_j$$

$T_n^* = T_n\left(X_1^*, X_2^*, \ldots X_n^*\right)$ l'estimateur de la médiane de X^* prend pour valeur $x_{(m)}^*$.

Notons N_j^* le nombre d'occurrences de $x_{(j)}$ lorsque l'on réalise $(X_1^*, X_2^*, \ldots X_n^*)$, la valeur prise par T_n^* est supérieure à $x_{(k)}$ si $\sum_{j=1}^{k} N_j^*$ est inférieur à m.

Quelle est la loi de N_j^* ?

C'est une binomiale de paramètres n et $p = \dfrac{1}{n}$

Quelle est la loi de $(N_1^*, N_2^*, \ldots N_n^*)$?

C'est une multinomiale de paramètres $\left(n, \dfrac{1}{n}, \dfrac{1}{n}, \ldots \dfrac{1}{n}\right)$

Quelle est la loi de $\sum_{j=1}^{k} N_j^*$ *?*

C'est une binomiale de paramètres n et $p = \dfrac{k}{n}$

Conclusion :

$$P\left(T_n^* > x_{(k)}\right) = P\left\{\sum_{j=1}^{k} N_j^* \leq m\right\} = \sum_{j=0}^{m}\left\{C_n^j \left(\frac{k}{n}\right)^j \left(\frac{n-k}{n}\right)^{n-j}\right\}$$

on en déduit :

$$\begin{aligned} p_k = P\left(T_n^* = x_{(k)}\right) &= P\left(T_n^* \leq x_{(k)}\right) - P\left(T_n^* \leq x_{(k-1)}\right) \\ &= \sum_{j=0}^{m}\left\{C_n^j \left(\frac{k-1}{n}\right)^j \left(\frac{n-k+1}{n}\right)^{n-j} - C_n^j \left(\frac{k}{n}\right)^j \left(\frac{n-k}{n}\right)^{n-j}\right\} \end{aligned}$$

$$\begin{aligned} E\left(T_n^*\right) &= \hat{\theta}_B = \textstyle\sum_k p_k x_{(k)} \\ \mathrm{Var}\left(T_n^*\right) &= \sum_k p_k x_{(k)}^2 - \left(\sum_k p_k x_{(k)}\right)^2 \end{aligned}$$

Exemple numérique : $n = 11$

k	1	2	3	4	5	6
p_k	0,00017	0,00703	0,04404	0,12150	0,20588	0,24274

k	7	8	9	10	11
p_k	0,20588	0,12150	0,04404	0,00703	0,00017

Considérons l'échantillon suivant, extrait d'une distribution uniforme sur $[0, 1]$:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
0,8734	0,8385	0,1705	0,4761	0,3223	0,6485	0,9906	0,1069	0,7261	0,3936	0,8269

ordonnée, nous obtenons :

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$
0,1069	0,1705	0,3223	0,3936	0,4761	0,6485	0,7261	0,8268	0,8385	0,8734	0,9906

$$\begin{aligned} \textstyle\sum_k p_k x_{(k)} &= 0,00017 * 0,1069 + 0,00703 * 0,1705 \\ &+ \cdots + 0,00017 * 0,9906 \end{aligned}$$

$$\hat{\theta}_B = 0,612$$

Remarque :

$\hat{\theta}_B$ est différent de $\hat{\theta}\left(\hat{\theta} = 0,6485\right)$.

On a :

$$\mathrm{Var}\left(T_n^*\right) = \sum_k p_k x_{(k)}^2 - \left(\sum_k p_k x_{(k)}^2\right)^2 = \sum_k p_k \left(x_k - \hat{\theta}_B\right)^2 = 0,016724$$

et :

$$\sum_k p_k \left(x_{(k)} - \hat{\theta}\right)^2 = 0,019210$$

b Variance de l'estimateur du coefficient de variation

L'exemple utilisé dans ce paragraphe met volontairement en jeu des échantillons de tailles réduites afin de permettre un meilleur suivi des calculs.

On considère un échantillon constitué de six couples de valeurs (x_i, y_i), la loi $\mathcal{F}$ du couple (X, Y) des variables parentes est inconnue :

x_i	$0,321$	$1,679$	$-2,188$	$0,889$	$-0,551$	$-0,188$
y_i	$-0,184$	$-0,647$	$-1,511$	$-0,851$	$-0,112$	$-1,152$

Le coefficient de corrélation a pour valeur $\hat{\rho} = 0,456$.

Comment estimer la variance de l'estimateur R de ρ ?

Ne connaissant pas la loi $\mathcal{F}$, on lui substitue la loi $\mathcal{F}^*$ associant à chaque couple (x_i, y_i) la même probabilité $\frac{1}{6}$.

$$P\left(X^* = x_i \text{ et } Y^* = y_i\right) = \frac{1}{6}$$

à l'estimateur R de ρ est substitué R^* :

$$R^* = \frac{\sum_i \left(X_i^* - \bar{X}^*\right)\left(Y_i^* - \bar{Y}^*\right)}{\left(\sum_i \left(X_i^* - \bar{X}^*\right)^2 \sum_i \left(Y_i^* - \bar{Y}^*\right)\right)^{1/2}}$$

et $\mathrm{Var}(R)$ est estimée par $\mathrm{Var}\left(R^*\right)$.

On pourrait étudier directement la loi de R^*, mais une telle approche peut paraître relativement complexe et l'on préfère opérer par simulation.

On effectue dix tirages $(B = 10)$ dans l'échantillon (x_i, y_i) :

Les résultats de ces tirages sont résumés dans le tableau présenté ci-dessous où l'on peut lire le nombre d'occurrences de chaque (x_i, y_i) dans les dix tirages et les estimations correspondantes de ρ.

x	y	1	2	3	4	5	6	7	8	9	10
$0,321$	$-0,184$	0	1	1	1	3	3	3	1	3	1
$1,679$	$-0,647$	1	1	1	0	0	1	0	1	0	0
$-2,188$	$-1,511$	1	0	1	1	0	1	1	2	0	1
$0,889$	$-0,851$	1	1	0	2	1	0	2	0	1	1
$-0,551$	$-0,112$	0	1	2	0	1	1	0	1	1	2
$-0,188$	$-1,152$	3	2	1	2	1	0	0	1	1	1

Pour chaque échantillon construit, on obtient une estimation de $\hat{\rho}_i$:

i	1	2	3	4	5	6	7	8	9	10
$\hat{\rho}_i$	$0,989$	$-0,035$	$0,420$	$0,676$	$-0,146$	$0,588$	$0,629$	$0,653$	$-0,146$	$0,427$

La moyenne des $\hat{\rho}_i$ est égale à 0,4087 et la variance :

$$\widehat{\mathrm{Var}}\,(R^*) \simeq 0,15$$

3.3 Taille de l'échantillon du Bootstrap

Le nombre B est la taille de l'échantillon "bootstrapé", c'est-à-dire le nombre de fois où l'on effectue des tirages dans la population constituée par l'échantillon représentant la loi $\mathcal{F}$.

Il est clair que lorsque $(n \to \infty)$, $\mathcal{F}^*$ approche $\mathcal{F}$ et que lorsque $(B \to \infty)$, les paramètres estimés à partir de l'échantillon de taille B et sous l'hypothèse de distribution $\mathcal{F}^*$, approchent leurs valeurs théoriques. La conjonction $(n \to \infty)$ et $(B \to \infty)$ valide la méthode du Bootstrap. Généralement, n est donné et le biométricien ne peut agir que sur B.

Le nombre d'échantillons de taille n que l'on peut construire à partir d'un n-échantillon $(x_1, x_2, \ldots x_n)$ est $C_{2n-1}^n = \dfrac{(2n-1)!}{n!(n-1)!}$

n	5	7	10
C_{2n-1}^n	126	2288	92378

Démonstration :

Soit C_n^p l'ensemble dont les éléments sont de la forme $(x_1^*, x_2^*, \ldots x_n^*)$, où $x_1^* \leq x_2^* \leq \cdots \leq x_n^*$ et $x_i^* \in (x_1, \ldots x_n)$. On note N_n^p la cardinalité de C_n^p.

Le nombre de fois où la valeur x_i se retrouve dans C_n^p est $\frac{p}{n}N_n^p$ et cela pour chaque $i = 1, \ldots, n$.

On définit $C_n^{p-1} = \left\{\left(x_2^*, \ldots, x_p^*\right) : \left(x_1, x_2^*, x_3^*, \ldots, x_p^*\right) \in C_n^p\right\}$. La cardinalité de C_n^{p-1} est N_n^{p-1} et le nombre de fois où la valeur x_1 se retrouve dans C_n^{p-1} est $\frac{p-1}{n}N_n^{p-1}$.

Puisque les seuls éléments de C_n^p comportant des valeurs x_1 sont de la forme $\left(x_1, x_2^*, \ldots, x_p^*\right)$ où $\left(x_2^*, \ldots, x_p^*\right) \in C_n^{p-1}$, on peut écrire :

$$\frac{p}{n}N_n^p = \frac{p-1}{n}N_n^{p-1} + N_n^p$$

et de là, la relation récursive :

$$\begin{aligned} N_n^p &= \frac{n+p-1}{p}N_n^{p-1} = \binom{n+p-1}{p}\binom{n+p-2}{p-1}\cdots\frac{n}{1} \\ &= C_{n+p-1}^p \end{aligned}$$

En posant $p = n$ on obtient le résultat recherché.

3.4 Comparaison Jackknife - Bootstrap

L'exemple décrit ci-dessous a été traité par Efron ; il est repris ici à partir de nouvelles données.

Soit (X, Y) un couple de valeurs aléatoires de lois normales :

$$\mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}; \sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

On dispose d'un échantillon de taille n :

$$(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$$

dont on peut extraire une estimation de ρ :

$$\hat{\rho} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2\right)^{1/2}}$$

et l'on désire estimer l'écart-type σ_ρ de l'estimateur de ρ.

On peut procéder selon trois méthodes :

on sait que sous l'hypothèse de normalité :

$$\hat{\sigma}_\rho \simeq \frac{1-\hat{\rho}^2}{\sqrt{n-3}}$$

on peut utiliser une approche Jackknife. On calcule les pseudo-valeurs $\tilde{\rho}_1\ \tilde{\rho}_2\ \cdots\ \tilde{\rho}_n$ et l'estimateur du Jackknife $\tilde{\rho}$:

$$\hat{\sigma}_J = \sqrt{\frac{1}{n(n-1)}\sum(\tilde{\rho}_i - \tilde{\rho})^2}$$

on peut mettre à contribution la technique du Bootstrap en effectuant B tirages à partir de l'échantillon :

$$(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$$

Pour chaque échantillon, on calcule $\hat{\rho}$, soit $\hat{\rho}_1, \hat{\rho}_2, \ldots \hat{\rho}_B$ puis une estimation de la moyenne $\bar{\rho}_B$ et de l'écart-type $\hat{\sigma}_B$:

$$\bar{\rho}_B = \frac{1}{B}\sum\hat{\rho}_i \qquad \hat{\sigma}_B = \left[\frac{1}{B-1}\sum(\hat{\rho}_i - \bar{\rho}_B)^2\right]^{1/2}$$

Pour comparer ces différentes méthodes on répète l'opération N fois. Les valeurs utilisées pour cette simulation sont :

$$N = 14000 \quad B = 128 \quad n = 8, 14, 20, \ldots 50$$

Les résultats sont présentés dans les tableaux suivants :

- q_1 : quartile inférieur
- q_3 : quartile supérieur
- Skewness : il s'agit d'un coefficient d'asymétrie qui est égal au moment centré d'ordre trois divisé par le cube de l'écart-type. Pour la loi normale, le coefficient est égal à zéro.
- Kurtosis : il s'agit d'un coefficient d'aplatissement qui est égal au moment centré d'ordre quatre divisé par le carré de la variance, la valeur trois étant soustraite à ce rapport. Pour la loi normale, le coefficient est égal à zéro.

Tableau I - Méthode : $\sigma = \frac{1-\rho^2}{\sqrt{n-3}}$

σ	$n=8$	$n=14$	$n=20$	$n=26$	$n=32$	$n=38$	$n=44$	$n=50$
Moyenne	0,3048	0,2163	0,1767	0,1527	0,1369	0,1247	0,1156	0,1081
Ecart − type	0,1072	0,0565	0,0388	0,0297	0,0238	0,0200	0,0173	0,0151
Skewness	−0,4488	−0,4309	−0,3760	−0,2910	−0,2940	−0,2491	−0,2320	−0,2215
Kurtosis	−0,8236	−0,5593	−0,4568	−0,4712	−0,3186	−0,2995	−0,2743	−0,2320
Maximum	0,447	0,302	0,243	0,209	0,186	0,169	0,156	0,146
q_3	0,401	0,263	0,207	0,175	0,155	0,139	0,128	0,119
Médiane	0,318	0,222	0,180	0,154	0,138	0,125	0,116	0,109
q_1	0,224	0,176	0,150	0,132	0,121	0,111	0,104	0,098
Minimum	0,008	0,030	0,036	0,057	0,048	0,052	0,052	0,051

Tableau II - Méthode Jackknife :

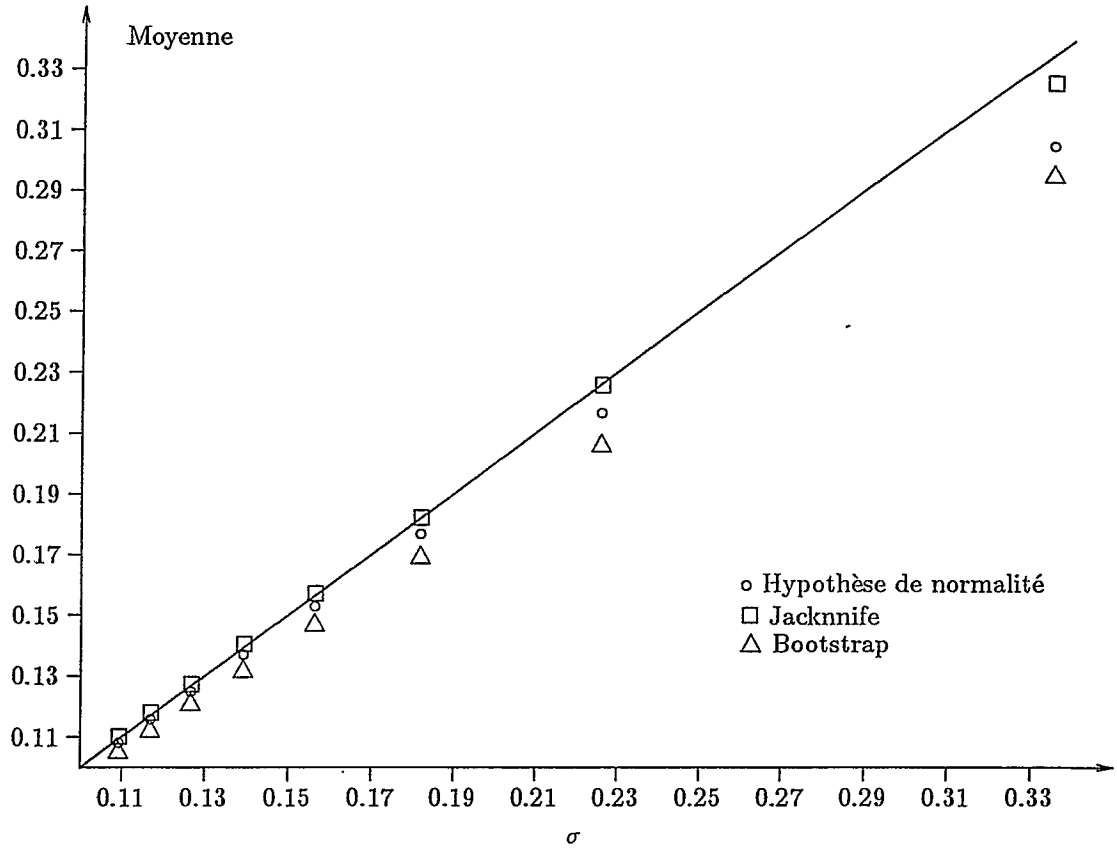
σ	$n = 8$	$n = 14$	$n = 20$	$n = 26$	$n = 32$	$n = 38$	$n = 44$	$n = 50$
Moyenne	0,3249	0,2246	0,1810	0,1560	0,1393	0,1263	0,1169	0,1090
Ecart − type	0,1680	0,0881	0,0589	0,0444	0,0353	0,0295	0,0255	0,0220
Skewness	0,7923	0,8470	0,6644	0,6567	0,6611	0,5223	0,5586	0,5000
Kurtosis	0,7950	1,5929	1,0479	1,0453	0,8154	0,7639	1,0139	0,8325
Maximum	1,261	0,776	0,571	0,446	0,333	0,301	0,321	0,244
q_3	0,425	0,275	0,217	0,182	0,161	0,145	0,133	0,103
Médiane	0,304	0,215	0,175	0,152	0,136	0,124	0,115	0,108
q_1	0,200	0,162	0,139	0,124	0,115	0,106	0,099	0,094
Minimum	0,006	0,020	0,037	0,041	0,041	0,042	0,033	0,046

Tableau III - Méthode Bootstrap :

σ	$n = 8$	$n = 14$	$n = 20$	$n = 26$	$n = 32$	$n = 38$	$n = 44$	$n = 50$
Moyenne	0,2949	0,2058	0,1691	0,1467	0,1316	0,1208	0,1121	0,1052
Ecart − type	0,1073	0,0653	0,0482	0,0375	0,0311	0,0267	0,0234	0,0208
Skewness	0,0159	0,2085	0,3182	0,2936	0,3130	0,3544	0,3810	0,3473
Kurtosis	−0,2558	−0,0529	0,1332	0,0637	0,0773	0,2171	0,2994	0,3046
Maximum	0,750	0,539	0,398	0,319	0,289	0,253	0,227	0,244
q_3	0,370	0,250	0,200	0,172	0,152	0,138	0,127	0,119
Médiane	0,298	0,205	0,167	0,145	0,130	0,119	0,111	0,104
q_1	0,219	0,160	0,135	0,120	0,109	0,102	0,096	0,091
Minimum	0,013	0,022	0,032	0,034	0,041	0,039	0,046	0,037

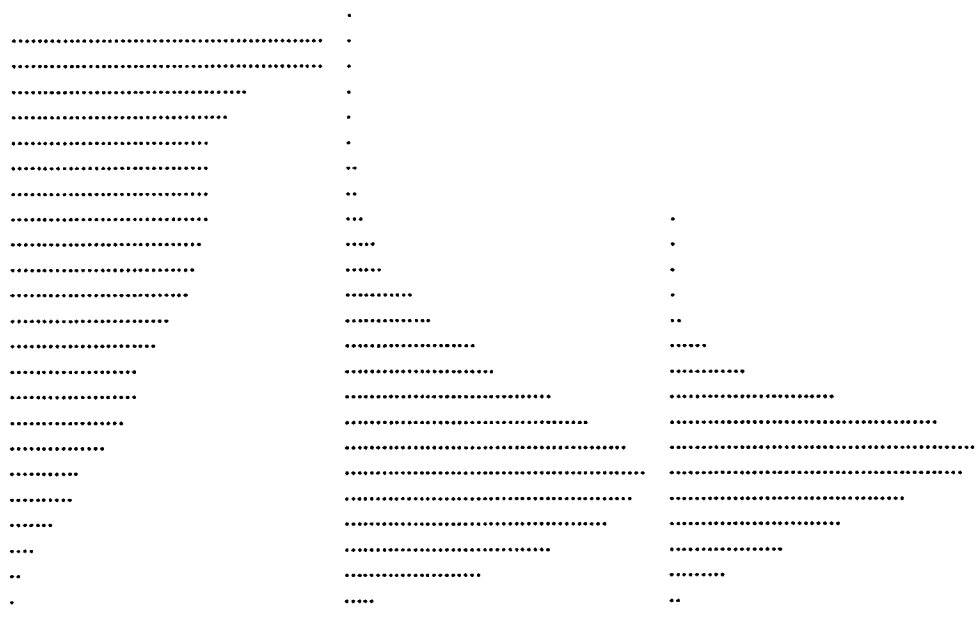
Conclusion :

a - Le Jackknife est l'estimateur qui minimise le biais, on peut le vérifier en plaçant sur un graphique les couples $(\sigma_{\text{calculé}}, \sigma_{\text{estimé}})$.



b - Le Bootstrap est moins dispersé que le Jackknife et tend plus rapidement vers une distribution du type normale :

$\boxed{n = 8}$



$n = 50$

Application à l'étude d'une régression linéaire

a Présentation du problème

Les données, mesurées par analyse thermique différentielle (ATD) sur des beurres laitiers français, sont constituées des pourcentages de glycérides liquides en fonction de la température, tous les 2°C, entre -5°C et 33°C. Elles permettent d'établir la courbe de fusion de chaque beurre étudié.

Cette courbe caractérise les propriétés thermiques de l'échantillon de beurre. Mais il est "lourd" de manipuler des courbes, d'où l'idée de résumer cette information sous la forme d'un modèle décrivant les variations de la courbe, dépendant de quelques paramètres aisément interprétables.

Avant d'aller plus loin, il faut savoir que les mesures ont été faites sur la matière grasse du beurre, constituée essentiellement de triglycérides. Chaque triglycéride est composé d'un glycérol et de trois acides gras. Et c'est la composition des acides gras dans les triglycérides qui détermine les propriétés physiques de la matière grasse, en particulier la courbe de fusion.

Nous nous intéressons ici à la courbe de fusion d'un beurre particulier, de Normandie, produit en saison hivernale. La courbe est présentée sur la figure 1. Et l'on obtient le profil de fusion en calculant les accroissements de taux de glycérides liquides en fonction de la température (figure 2).

Figure 1 : taux de glycérides liquides en fonction de la température

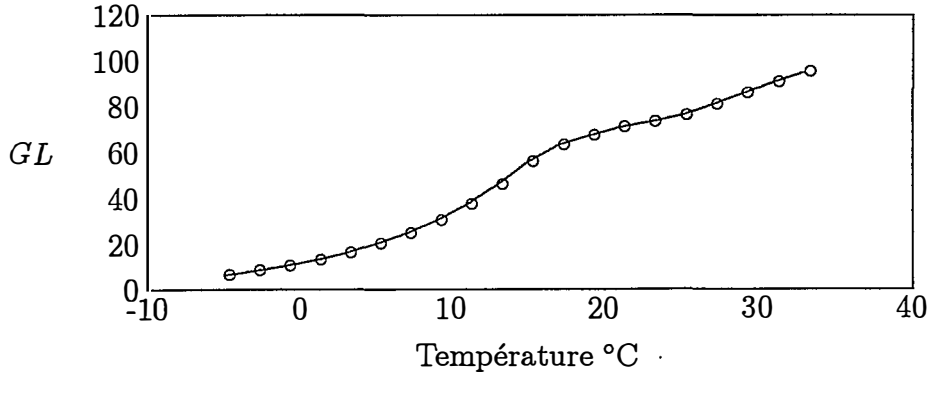
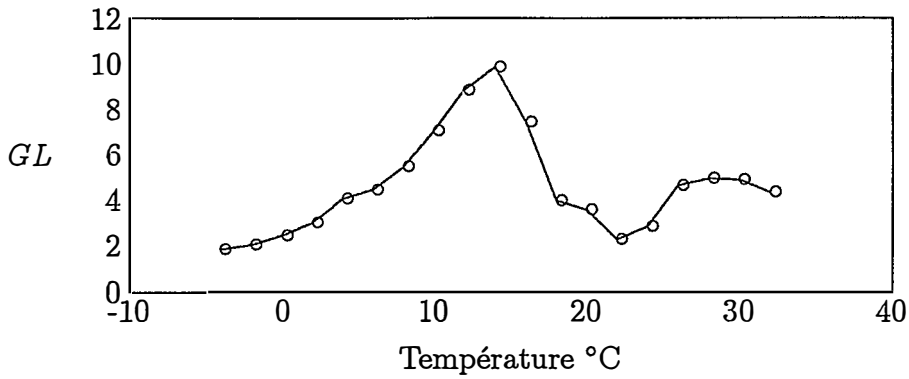


Figure 2 : profil de fusion des glycérides



Après avoir défini le modèle non linéaire donnant la courbe de fusion, notre objectif est de valider ou d'améliorer l'estimation des paramètres (et leur distribution) obtenue par la méthode des moindres carrés. C'est ce qui nous a amené à utiliser des méthodes de rééchantillonnage telles qu'Eustache (ou Jackknife) et Cyrano (ou Bootstrap).

b Modèle utilisé

Il apparaît clairement qu'il y a deux pics bien différenciés. Deux masses de glycérides fondent successivement lors de la fusion. Alors l'idée est d'ajuster un modèle correspondant à la somme de deux fonctions logistiques : une fonction logistique généralisée pour le premier pic, qui est asymétrique, et une fonction logistique simple pour le deuxième pic.

On sait que le pourcentage de glycérides liquides (noté GL par la suite) tend vers 0 pour les températures basses, et vers 100 pour les températures élevées. Ceci impose la valeur des deux asymptotes. Le modèle somme de deux logistiques suivant a été utilisé :

$$GL = \frac{A}{(1 + F\mathrm{e}^{B(C-t)})^{\frac{1}{F}}} + \frac{(100 - A)}{(1 + \mathrm{e}^{D(E-t)})}$$

où les six paramètres sont aisément interprétables :
A est la proportion (en %) de glycérides contenus dans le premier pic de fusion,
B et D mesurent, respectivement, l'étroitesse du premier et du deuxième pics de fusion,
C et E correspondent, respectivement, à la température (en °C) du premier et du deuxième pics de fusion,
et F mesure l'asymétrie du premier pic de fusion ($F = 1$ si le pic est symétrique).

Les ajustements ont été réalisés sur les données d'une série de beurres. On aboutit à des déterminants des matrices X^tX (matrices d'information de Fisher au minimum des sommes de carrés d'écarts) très petits (10^{-5} à 10^{-8}) et à des corrélations élevées

entre les paramètres (surtout entre C et F : 0,995 environ). Ceci a pour conséquence une grande imprécision sur les paramètres, et traduit une surparamétrisation du modèle.

Le choix d'une valeur fixe pour l'asymétrie a permis de pallier cet inconvénient. Une valeur médiane a été prise : $F = 4$. D'où le modèle finalement retenu :

$$GL = \frac{A}{\left(1 + 4\mathrm{e}^{B(C-t)}\right)^{0.25}} + \frac{(100 - A)}{\left(1 + \mathrm{e}^{D(E-t)}\right)}$$

les résultats sont alors plus satisfaisants.

Dans le cas du beurre présenté ci-dessus, avec la méthode des moindres carrés, la somme des carrés résiduels est de 3,613 (on a 20 observations et 5 paramètres estimés d'où 15 degrés de liberté, l'écart-type résiduel est donc de 0,491), le coefficient de détermination R^2 est de 0,99994.

	paramètres estimés	écarts − type
A	$72,06$	$0,70$
B	$0,4513$	$5,810^{-3}$
C	$13,45$	$0,14$
D	$0,3939$	$2,5110^{-2}$
E	$29,07$	$0,19$

estimation de la matrice de corrélation des paramètres :

A	$1,000$				
B	$-0,543$	$1,000$			
C	$0,919$	$-0,636$	$1,000$		
D	$0,784$	$-0,318$	$0,668$	$1,000$	
E	$0,855$	$-0,433$	$0,773$	$0,651$	$1,000$

valeurs observées, ajustées et résidus (en %) :

t	GLobs	GLest	résidus
$-5°C$	6,35	6,357	$-0,007$
$-3°C$	8,25	7,966	0,284
$-1°C$	10,35	9,982	0,368
$1°C$	12,85	12,508	0,342
$3°C$	15,90	15,670	0,230
$5°C$	20,00	19,621	0.379
$7°C$	24,50	24,541	$-0,041$
$9°C$	30,00	30,608	$-0,608$
$11°C$	37,05	37,922	$-0,872$
$13°C$	45,90	46,298	$-0,398$
$15°C$	55,75	54,934	0,816
$17°C$	63,20	62,410	0,790
$19°C$	67,25	67,667	$-0,417$
$21°C$	70,90	70,973	$-0,073$
$23°C$	73,25	73,470	$-0,220$
$25°C$	76,15	76,360	$-0,210$
$27°C$	80,85	80,481	0,369
$29°C$	85,85	85,784	0,066
$31°C$	90,80	91,082	$-0,282$
$33°C$	95,20	95,096	0,104

La lecture des résidus montre que les hypothèses faites sur le terme d'erreur résiduelle ne sont pas correctement vérifiées. Les variables aléatoires E_i ne sont pas indépendantes. Le modèle est faux en toute rigueur.

Mais il ne sert à rien de vouloir aller plus loin dans l'ajustement. Notre problème n'est pas de faire de la prédiction du taux de glycérides liquides à température donnée, ce qui nécessiterait de tenir compte des autocorrélations. Il s'agit ici de caractériser les propriétés thermiques des échantillons de beurre, donc de faire une estimation fiable des paramètres qui permettent cette caractérisation. Le modèle sert donc de filtre pour résumer l'information contenue dans la courbe $GL = f(t)$. D'autre part, les résidus sont faibles en valeur absolue.

Une fois ce modèle choisi, il était intéressant de valider ou d'améliorer l'estimation de ces paramètres par des méthodes de rééchantillonnage comme Eustache et Cyrano. Et dans le cas où ces paramètres ne sont pas trop instables, les ajustements devaient être faits sur l'ensemble des beurres étudiés par ailleurs.

c Estimation des paramètres par les méthodes de rééchantillonnage

Méthode Eustache (Jackknife)

On considère θ un estimateur d'un paramètre μ basé sur un échantillon $(X_1, \ldots, X_n)$; θ, μ et $X_i (i = 1, \ldots n)$ pouvant être des vecteurs. On définit la pseudo-valeur :

$$\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$$

où $\hat{\theta}_{(-i)}$ est l'estimateur θ du paramètre μ, calculé sur l'échantillon $(X_1, \ldots, , X_n)$ privé de l'observation X_i. On définit alors l'estimateur d'Eustache :

$$\tilde{\theta} = \frac{1}{n}\tilde{\theta}_{(i)}$$

$\tilde{\theta}$ a une distribution asymptotiquement normale, dont l'estimation de la matrice de variance-covariance est donnée par :

$$\frac{\left(\sum_i \left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)\left(\tilde{\theta}_{(i)} - \tilde{\theta}\right)'\right)}{n(n-1)}$$

Les pseudo-valeurs $\tilde{\theta}_{(i)}$ permettent aussi de détecter les valeurs suspectes et d'évaluer leur influence sur les estimateurs.

Cette méthode a été appliquée à nos données d'ATD. Les X_i définis ci-dessus correspondent ici aux couples $(t_i; GL_i)\, i = 1 \ldots 20$, les t_i étant les 20 températures (de -5°C à 33°C) et GL_i les taux de glycérides liquides mesurés.

Les vecteurs μ et θ correspondent respectivement aux cinq paramètres A, B, C, D et E et à leurs estimateurs.

D'où les estimations d'Eustache suivantes :

	paramètres estimés	écarts − type
A	$71,94$	$0,57$
B	$0,4524$	$6,710^{-3}$
C	$13,42$	$0,16$
D	$0,3891$	$1,9810^{-2}$
E	$29,04$	$0,16$

et l'estimation de la matrice de corrélation des paramètres :

A	$1,000$				
B	$-0,398$	$1,000$			
C	$0,811$	$-0,719$	$1,000$		
D	$0,789$	$-0,143$	$0,513$	$1,000$	
E	$0,876$	$-0,279$	$0,659$	$0,740$	$1,000$

La conclusion qui s'impose est que ces estimations d'Eustache sont très proches de celles des moindres carrés, que ce soit pour les estimations des paramètres ou pour l'évaluation de leur dispersion. On a une bonne fiabilité dans l'estimation des paramètres par les moindres carrés.

Une autre utilité de cette technique est de voir s'il y a des données suspectes. L'étude des histogrammes des pseudo-valeurs montre que quelques-unes sont un peu écartées des autres.

Ce n'est pas le cas pour le paramètre A, mais les pseudo-valeurs du paramètre B à 15 et 17°C et de C à 15°C s'écartent un peu des autres. Elles correspondent à des résidus élevés dans l'ajustement simple vu ci-dessus. Quant à D et E, leur pseudo-valeur à 27°C est peu élevée, sans être véritablement suspecte.

Méthode Cyrano (Bootstrap)

On se place dans le même cadre que pour la méthode Eustache : on considère θ un estimateur d'un paramètre μ basé sur un échantillon $(X_1, \ldots, X_n)$; θ, μ et X_i $(i = 1, \ldots, n)$ pouvant être des vecteurs.

Cette technique, comme Eustache, est basée sur la génération d'échantillons artificiels à partir des n observations. On génère N échantillons Cyrano, de n individus chacun. Chaque échantillon est obtenu par tirage aléatoire simple avec remise de n individus parmi les n observations de l'échantillon de départ.

Soit $\hat{\theta}_{B(I)}$ l'estimateur de μ basé sur l'échantillon Cyrano I $(I = 1, \ldots, N)$. On définit l'estimateur Cyrano :

$$\hat{\theta}_B = \frac{1}{N} \sum_{I=1}^{N} \hat{\theta}_{B(I)}$$

dont on peut estimer la matrice de variance-covariance par :

$$\frac{\left(\sum_I \left(\hat{\theta}_{B(I)} - \hat{\theta}_B\right) \left(\hat{\theta}_{B(I)} - \hat{\theta}_B\right)'\right)}{(N-1)}$$

La distribution des $\hat{\theta}_{B(I)}$ estime celle de l'estimateur θ.

$N = 100$ échantillons Cyrano ont été tirés, on obtient alors les estimations suivantes des paramètres et de leur dispersion :

	paramètres estimés	écarts − type
A	$72,40$	$0,62$
B	$0,4514$	$6,310^{-3}$
C	$13,46$	$0,15$
D	$0,4163$	$9,2710^{-2}$
E	$29,04$	$0,33$

et l'estimation de la matrice de corrélation des paramètres :

A	$1,000$				
B	$-0,376$	$1,000$			
C	$0,848$	$-0,593$	$1,000$		
D	$0,483$	$-0,036$	$0,275$	$1,000$	
E	$0,195$	$-0,147$	$0,206$	$-0,605$	$1,000$

On constate que, pour les paramètres caractérisant le premier pic (A, B et$C$), les estimations Cyrano sont en tout point identiques à celles des moindres carrés et à celles d'Eustache. Ce qui constitue une confirmation de la fiabilité de l'estimation de ces paramètres.

L'estimation des paramètres caractérisant le deuxième pic (D et E) n'est pas non plus sensiblement différente des estimations précédentes. Par contre l'estimation de leur dispersion est différente. L'écart-type du paramètre D (étroitesse du deuxième pic) est cinq fois plus grand, celui du paramètre E (température du deuxième pic) l'est deux fois plus. Et les corrélations entre ces deux paramètres d'une part, avec les trois autres d'autre part, sont nettement différentes.

L'étude de la distribution des paramètres D et E estimée par Cyrano donne la raison de ces différences. On peut voir des valeurs franchement suspectes.

Ceci illustre la limitation de la méthode Cyrano dans le cas où le nombre d'observations n est faible. Les valeurs suspectes correspondent à des échantillons Cyrano pour lesquels les paramètres D et E (du deuxième pic) sont estimés avec très peu de précision car les points de la dernière partie de la courbe sont absents par le fait du hasard des tirages.

On peut dire que si les 4 derniers points (de 27 à 33°C) sont absents de l'échantillon tiré, on ne peut estimer correctement D et E. Or, la probabilité de tirer un tel échantillon est égale à la probabilité de tirer à 20 reprises des points parmi les 16 premiers (de -5 à 25°C), d'où :

$$p = (0,8)^{20} = 0,0115$$

comme $N = 100$ tirages d'échantillons sont effectués, la probabilité d'obtenir un tel échantillon est égale à :

$$1 - \left(1 - (0,8)^{20}\right)^{100} = 0,686$$

Donc les échantillons tirés où manquent les 4 derniers points, ou bien 3 sur les 4, expliquent les valeurs suspectes des paramètres D et E. Ce qui amène à penser que la méthode Cyrano ne permet pas l'estimation de la distribution des paramètres dont le calcul est basé sur trop peu de données. Ici le calcul de D et E est basé sur 4 observations seulement, peut-être 5.

d Conclusion

Le bilan de ces études de rééchantillonnage pour l'estimation des 5 paramètres du modèle décrivant l'évolution du taux de glycérides liquides en fonction de la température est double. L'estimation de ces paramètres par les moindres carrés est fiable et n'introduit pas de biais important, pour le beurre étudié. Et l'estimation de la dispersion de ces paramètres est également fiable. Ce qui justifiait de faire ces ajustements non linéaires en routine pour caractériser les beurres.

Mais on peut également tirer un enseignement de cette étude sur les méthodes elles-mêmes. Autant Eustache ne pose pas de problème, autant Cyrano, tel qu'il est appliqué ici, n'est pas très intéressant. Aussi il a été proposé, pour règler ce problème, de tirer aléatoirement les seuls résidus, en régression non linéaire (et plus généralement en expérimentation planifiée), plutôt que les données comme c'est le cas habituellement.

Etude de la stabilité de l'ACP par le Bootstrap

5.1 Introduction

Les observations étant toujours sujettes aux erreurs de mesure ou d'échantillonnage, les résultats d'une Analyse en Composantes Principales (ACP) peuvent varier d'un échantillon à l'autre.

La méthode de rééchantillonnage du Bootstrap permet, grâce à la simulation de plusieurs échantillons, d'estimer la variabilité des résultats de l'ACP et donc d'estimer leur fiabilité.

5.2 Rappels

a But de l'ACP

Le but de l'ACP est de rechercher une approximation de la matrice de données initiale $X(n,p)$, à n individus et p variables mesurées sur chaque individu, par une matrice de rang inférieur q. Si les n lignes de X sont considérées comme les coordonnées de n points dans un espace à p dimensions, on peut alors représenter graphiquement X dans un sous-espace de plus faible dimension q.

Donc, le problème qui se pose concerne le choix du nombre de composantes (ou de dimensions) qui doivent être retenues. Les règles souvent utilisées prenent en compte deux types de critères :

- la part de variance expliquée (ou taux d'inertie),
- le comportement en "éboulis" des valeurs propres.

b Principe de l'ACP

L'ajustement progressif de sous-espaces "emboîtés" de dimensions k croissantes ($k = 1, \ldots, p$) au nuage de n points, par la méthode des moindres carrés, fournit les p vecteurs propres (μ_k) correspondant aux p valeurs propres (λ_k) de la matrice symétrique $X'X$, avec ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$). Les coordonnées Z_k des points individus sur l'axe k sont alors les produits scalaires constituant les lignes de Xu_k.

On écrit :

$$\underset{(p,n)}{X'} \quad \underset{(n,p)}{X} \quad \underset{(p,p)}{U} \quad = \quad \underset{(p,p)}{U} \quad \underset{(p,p)}{\Lambda}$$

et l'on obtient la matrice des variables transformées par l'ACP :

$$\underset{(n,p)}{Z} \quad = \quad \underset{(n,p)}{X} \quad \underset{(p,p)}{U}$$

On note :

$$\begin{cases} V & = X'X \text{ la matrice de variance} - \text{covariance de } X \\ \Delta & \text{diag}(V) \text{ la matrice diagonale des variances de } X \\ \Lambda & \text{la matrice diagonale des valeurs propres de } X' \\ U & \text{la matrice orthogonale des vecteurs propres de } X'X \\ R & \text{la matrice des corrélations de } X \end{cases}$$

5.3 Application du Bootstrap à l'ACP

a Principe du Bootstrap

La méthode du Bootstrap consiste à réaliser un tirage avec remise de B n-sous-échantillons à partir d'un n-échantillon lui-même représentatif d'une population d'origine (P). Si l'on s'intéresse à une statistique T, la distribution F de T du n-échantillon dans la population d'origine P est simulée par la distribution $F^{(b)}$ de la même statistique $T^{(b)}(b = 1, 2, \ldots, B)$ calculée dans chaque sous-échantillon. Cette simulation donne accès aux caractéristiques de $F^{(b)}$ (la moyenne, le biais, la variance, ...) qui sont les estimations des caractéristiques de F.

b Tirage des sous-échantillons

Le $b^{\text{ème}}$ n-sous-échantillon est obtenu par tirage équiprobable, avec remise, de n unités dans le n-échantillon.

Soit X la matrice des observations dans l'échantillon, soit $P^{(b)}$ une matrice des pondérations associées au $b^{\text{ème}}$ sous-échantillon, de terme général p_{ij}, on peut noter la matrice des observations du $b^{\text{ème}}$ sous-échantillon :

$$\underset{(n,p)}{X^{(b)}} \quad = \quad \underset{(n,n)}{P^{(b)}} \quad \underset{(n,p)}{X}$$

Exemple :

Prenons $n = 4$ et admettons que l'on tire :

- 2 fois la $1^{\text{ère}}$ ligne d'observations dans X

- 1 fois la $2^{\text{ème}}$ ligne d'observations dans X
- 0 fois la $3^{\text{ème}}$ ligne d'observations dans X
- 1 fois la $4^{\text{ème}}$ ligne d'observations dans X

On aura :

$$P^{(b)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

avec :

$$\begin{cases} \sum_i p_{ij} & = \text{ nombre d'occurrences de la ligne } j \\ \sum_j p_{ij} & = 1 \\ \sum_{ij} p_{ij} & = n \end{cases}$$

Le Bootstrap peut-être réalisé pour l'ACP non-normée (diagonalisation de la matrice de variance-covariance V) ou normée (diagonalisation de la matrice des corrélations R).

Trois natures de tirages sont envisageables pour chaque type d'ACP :

- des tirages sur les données brutes,
- des tirages sur les données centrées,
- des tirages sur les variables transformées.

Dans tous les cas, les données "bootstrapées" constituant chaque sous-échantillon s'écriront sous la forme du produit d'une matrice de pondération par la matrice de données associée à l'échantillon considéré.

c Notations

Des exposants permettent de signifier l'origine des paramètres :

- paramètres associés à la population de base : exposant (0),
- paramètres associés au n-échantillon : pas d'exposant,
- paramètres associés aux n-sous-échantillons : exposant (b).

Pour préciser la nature des données "bootstrapées" et le type de l'ACP étudié, les paramètres sont indicés :

- ACP non-normée : indice (V)
- ACP normée : indice (R)

5.4 Evaluation de la stabilité de l'ACP

• Pour l'ACP non-normée, les paramètres intéressants sont :

$$V^{(0)} \quad , \quad \Lambda_V^{(0)} \text{ et } U_V^{(0)}$$

leurs estimations sur l'échantillon sont respectivement :

$$V \quad , \quad \Lambda_V \text{ et } U_V$$

la stabilité de l'ACP est mesurée par la variabilité de ces paramètres :

$$\mathrm{E}\left(V - V^{(0)}\right)^2 \quad , \quad \mathrm{E}\left(\Lambda_V - \Lambda_V^{(0)}\right)^2 \text{ et } \mathrm{E}\left(U_V - U_V^{(0)}\right)^2$$

la méthode du Bootstrap fournit une estimation de cette variabilité par :

$$\mathrm{E}_B\left(V^{(b)} - V\right)^2 \quad , \quad \mathrm{E}\left(\Lambda_V^{(b)} - \Lambda_V\right)^2 \text{ et } \mathrm{E}_B\left(U_V^{(b)} - U_V\right)^2$$

où E_B est l'espérance calculée sur l'ensemble des B sous-échantillons.

• Pour l'ACP normée, les paramètres intéressants sont :

$$R^{(0)} \quad , \quad \Lambda_R^{(0)} \text{ et } U_R^{(0)}$$

la stabilité de l'ACP est mesurée par :

$$\mathrm{E}\left(R - R^{(0)}\right)^2 \quad , \quad \mathrm{E}\left(\Lambda_R - \Lambda_R^{(0)}\right)^2 \text{ et } \mathrm{E}\left(U_R - U_R^{(0)}\right)^2$$

et estimée par :

$$\mathrm{E}_B\left(R^{(b)} - R\right)^2 \quad , \quad \mathrm{E}_B\left(\Lambda_R^{(b)} - \Lambda_R\right)^2 \text{ et } \mathrm{E}_B\left(U_R^{(b)} - U_R\right)^2$$

Quel que soit le type d'ACP, afin d'attribuer un scalaire à chacune de ces mesures, on "déroule" chaque matrice (p, p) en un vecteur $(p^2, 1)$ avant de calculer les espérances.

a Etude de la stabilité de l'ACP non-normée

La méthode du Bootstrap peut être appliquée indifféremment aux données brutes (X) ou aux données centrées $(Y = X - \mu)$. En effet, la variance étant invariante par translation :

$$V_Y = V_X \text{ et } V_Y^{(b)} = V_X^{(b)}$$

Donc $\mathrm{E}_B\left(V_Y^{(b)} - V_Y\right)^2$ peut être utilisée pour estimer $\mathrm{E}\left(V_X - V_X^{(0)}\right)^2$.

Une troisième possibilité consiste en un tirage de B sous-échantillons parmi les variables transformées : Z

$$\underset{(n,p)}{Z} = \underset{(n,p)}{X} \; \underset{(p,p)}{U_{XV}}$$

En fait, cette procédure est équivalente aux précédentes. En effet :

- $$\left.\begin{aligned} V_Z^{(b)} &= U'_{XV}\, V_X^{(b)}\, U_{XV} \\ V_Z &= U'_{XV}\, V_X\, U_{XV} \end{aligned}\right\} \quad \text{donc} \quad V_X^{(b)} - V_X = U_{XV}\left(V_Z^{(b)} - V_Z\right) U'_{XV}$$

et par conséquent $\mathrm{E}_B\left(U_{XV}\left(V_Z^{(b)} - V_Z\right) U'_{XV}\right)^2$ estime $\mathrm{E}\left(V_X - V_X^{(0)}\right)^2$

- $$\left.\begin{aligned} V_Z^{(b)} &= U'_{XV}\, V_X^{(b)}\, U_{XV} \\ V_X^{(b)} &= U_{XV}^{(b)}\, \Lambda_X^{(b)}\, U_{XV}^{\prime(b)} \end{aligned}\right\} \quad \text{donc} \quad V_Z^{(b)} = U'_{XV}\, U_{XV}^{(b)}\, \Lambda_X^{(b)}\, U_{XV}^{\prime(b)}\, U_{XV}$$

avec :

$$V_Z^{(b)} = U_{ZV}^{(b)}\, \Lambda_Z^{(b)}\, U_{ZV}^{\prime(b)}$$

on obtient :

$$\left.\begin{aligned} \Lambda_X^{(b)} &= \Lambda_Z^{(b)} \\ U_{ZV}^{\prime(b)} &= U_{XV}^{\prime(b)}\, U_{XV} \end{aligned}\right\} \quad \text{donc} \quad \begin{cases} \mathrm{E}_B\left(\Lambda_Z^{(b)} - \Lambda_X\right)^2 & \text{estime} \quad \mathrm{E}\left(\Lambda_V - \Lambda_X^{(0)}\right)^2 \\ \mathrm{E}\left(U_{ZV}^{\prime(b)} U'_{XV} - U'_{XV}\right)^2 & \text{estime} \quad \mathrm{E}\left(U'_{XV} - U_{XV}^{\prime(0)}\right)^2 \end{cases}$$

Travailler sur Z présente deux intérêts :

- la différence entre $V_Z^{(b)}$ et V_Z est facile à visualiser car V_Z est diagonale (en effet, V_Z n'est autre que Λ_{XV}),
- la diagonalisation de $V_Z^{(b)}$ est plus économique en temps de calcul.

De plus, lorsque le nombre de variables est élevé, la seule prise en compte des k premières variables de l'ACP peut permettre de réduire davantage le temps de calcul tout en fournissant une bonne approximation de V_X et $V_X^{(b)}$.

b Etude de la stabilité de l'ACP normée

Dans le cas de l'ACP normée, la normalisation fournie par $\Delta_X^{-\frac{1}{2}}$ est différente pour chaque sous-échantillon. La matrice des corrélations s'écrit :

$$R_X^{(b)} = \Delta_X^{(b)-\frac{1}{2}}\, V_X^{(b)}\, \Delta_X^{(b)-\frac{1}{2}}$$

De même que lors de l'étude de la stabilité de l'ACP non-normée, on peut indifféremment utiliser X, Y, ou W (avec $W = Y\Delta_X^{-\frac{1}{2}}$, la matrice des données centrées réduites) dans le calcul de $R_X^{(b)}$.

On appelle [A] la méthode du Bootstrap qui consiste à mesurer la stabilité des estimations à partir de $R_X^{(b)}$ (ou $R_Y^{(b)}$ ou $R_W^{(b)}$).

Par contre, à l'inverse des résultats obtenus pour l'ACP non-normée, il devient sans intérêt d'utiliser la matrice des corrélations $R_S^{(b)}$ des variables transformées par l'ACP dans les sous-échantillons : S

$$\underset{(n,p)}{S} = \underset{(n,p)}{W} \; \underset{(p,p)}{U_{XR}}$$

En effet, la matrice des corrélations correspondantes s'écrit :

$$R_S^{(b)} = \left[\text{diag}\left(U'_{XR}\, V_W^{(b)}\, U_{XR}\right)\right]^{-\frac{1}{2}} U'_{XR}\, V_W^{(b)}\, U_{XR} \left[\text{diag}\left(U_{XR}\, V_W^{(b)}\, U'_{XR}\right)\right]^{-\frac{1}{2}}$$

Or, il est impossible de retrouver $R_X^{(b)}$ à partir de $R_S^{(b)}$ et donc d'estimer la stabilité de l'ACP normée en appliquant la méthode du Bootstrap aux variables transformées. Toutefois, à défaut de pouvoir estimer $R_X^{(b)}$, on peut s'en approcher en estimant $V_W^{(b)}$.

On appelle [B] la méthode du Bootstrap qui consiste à mesurer la stabilité des estimations à partir de $V_W^{(b)}$. *Cette approximation se justifie par le fait qu'il est raisonnable de penser que le comportement de* $R_X - R_X^{(0)}$ *n'est pas trop éloigné de celui de* $V_W - R_X^{(0)}$.

$$\left.\begin{aligned} V_W^{(b)} &= U_{XR}\, V_S^{(b)}\, U'_{XR} \\ R_X &= U_{XR}\, \Lambda_{XR}\, U'_{XR} \end{aligned}\right\} \quad \text{donc} \quad V_W^{(b)} - R_X = U_{XR}\left(V_S^{(b)} - \Lambda_{XR}\right) U'_{XR}$$

et par conséquent :

$$\text{E}_B\left(U_{XR}\left(V_S^{(b)} - \Lambda_{XR}\right) U'_{XR}\right)^2 \;\; \text{estime} \;\; \text{E}\left(V_W - R_X^{(0)}\right)^2$$

5.5 Mesures de la stabilité de l'ACP

Afin de calculer un nombre raisonnable de paramètres, il est préférable de se concentrer sur les plus intéressants. En Analyse en Composantes Principales, le problème est de pouvoir déterminer la dimension de l'espace de représentation optimum. Il s'agit de conserver toutes les caractéristiques stables et importantes des données étudiées tout en ignorant les axes instables et sans signification.

Dans cet esprit, on peut envisager de mesurer la stabilité de deux types de paramètres :

- les valeurs propres,
- les sous-espaces de représentation.

a Stabilité des valeurs propres

Soit Λ, la matrice diagonale des valeurs propres ordonnées $(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p)$, pour chaque axe k, on calcule la moyenne $\left(\bar{\lambda}_k\right)$, l'écart-type $(S_{\lambda k})$ et le biais associé à la valeur propre λ_k correspondante.

On écrit :

$$\begin{cases} \bar{\lambda}_k & = \dfrac{1}{B}\sum_b \lambda_k^{(b)} \\ S_{\lambda k} & = \sqrt{\dfrac{1}{B}\sum_b \left(\lambda_k^{(b)} - \bar{\lambda}_k\right)^2} \\ \text{Biais}_{\lambda k} & = \mathrm{E}_B\left(\lambda_k^{(b)} - \lambda_k\right) = \bar{\lambda}_k - \lambda_k \end{cases}$$

On démontre que le biais associé à la première valeur propre est positif, tandis que le biais associé à la dernière valeur propre est négatif. En d'autres termes, la proportion de variabilité expliquée par le premier axe de l'ACP sur un sous-échantillon ne peut être que supérieure ou égale à la proportion de variabilité expliquée par le premier axe de l'ACP réalisée sur l'échantillon d'origine.

Par construction, le raisonnement inverse doit être appliqué à la proportion de variabilité expliquée par le dernier axe de l'ACP appliquée à un sous-échantillon de l'échantillon d'origine.

Soit $Z_1, \ldots, Z_p$ les variables de l'ACP, on obtient par conséquent :

$$\begin{cases} \mathrm{Var}\left(Z_1^{(0)}\right) \leq \mathrm{Var}\left(Z_1\right) = \lambda_1 \quad \text{donc} \quad \mathrm{E}_B\left(\lambda_1^{(b)} - \lambda_1\right) \geq 0 \\ \mathrm{Var}\left(Z_p^{(0)}\right) \geq \mathrm{Var}\left(Z_p\right) = \lambda_p \quad \text{donc} \quad \mathrm{E}_B\left(\lambda_p^{(b)} - \lambda_p\right) \leq 0 \end{cases}$$

b Stabilité des sous-espaces de représentation

Soit E_k, l'espérance engendrée par les k premiers vecteurs propres, on mesure la variabilité de E_k autour de $\mathrm{E}_k^{(0)}$ par la mesure angulaire entre sous-espaces utilisée en analyse canonique :

$$A_k = \sum_{i,j \leq k} \rho_{ij}^2 \quad \text{avec} \quad \rho_{ij} = \mathrm{Corr}\left(Z_i^{(0)}, Z_j\right)$$

où ρ_{ij} est la corrélation entre la $i^{\text{ème}}$ variable de l'ACP sur la population et la $j^{\text{ème}}$ variable de l'ACP sur l'échantillon.

Remarque : la stabilité parfaite de tous les $k^{\text{èmes}}$ axes est caractérisée par :

$$\mathrm{E}\left(\rho_{kk}^2\right) = 1 \quad \text{pour } k = 1, \ldots, p$$

ce qui revient à écrire :

$$\mathrm{E}\left(A_k\right) = k \qquad \forall k$$

Pour chaque axe k, on calcule la moyenne $\bar{A}_k$, l'écart-type S_{Ak} et l'écart-type quadratique moyen EQM_k par rapport à k.

On écrit :

$$\begin{cases} \bar{A}_k & = \dfrac{1}{B} \displaystyle\sum_b A_k^{(b)} \\ S_{Ak} & = \sqrt{\dfrac{1}{B} \displaystyle\sum_b \left(A_k^{(b)} - \bar{A}_k\right)^2} \\ \mathrm{EQM}_k & = \mathrm{E}_B \left(A_k^{(b)} - k\right)^2 \end{cases}$$

Les $A_k^{(b)}$ sont obtenus à partir des $\rho_{ij}^{(b)}$ mesurés de la façon suivante :

$$\left.\begin{array}{lll} \bullet & Z_j^{(b)} & = X\ U_j^{(b)} \\ & \mathrm{Var}\left(Z_j^{(b)}\right) & = U_j'^{(b)}\ V\ U_j^{(b)} \\ \bullet & Z_i & = X\ U_i \\ & \mathrm{Var}\left(Z_i\right) & = U_i'\ V\ U_i = \lambda_i \\ \bullet & Z_j'^{(b)} Z_i & = U_j'^{(b)}\ V\ U_i \end{array}\right\} \quad \text{donc} \quad \begin{array}{l} \rho_{ij} = \left(U_j'^{(b)}\ V\ U_i\right) \lambda_i^{-\frac{1}{2}} \left(U_j'^{(b)}\ V\ U_j^{(b)}\right)^{-\frac{1}{2}} \\ \qquad\quad = \left(\lambda_i^{\frac{1}{2}}\ U_j'^{(b)}\ U_i\right) \left(U_j'^{(b)}\ V\ U_j^{(b)}\right)^{-\frac{1}{2}} \end{array}$$

où V est la matrice de corrélation R_X ou la matrice de variance-covariance V_X , selon que l'ACP est normée ou non. U_i est le $i^{\text{ème}}$ vecteur propre de V.

5.6 Exemples

La stabilité de l'ACP normée est ici étudiée sur deux séries de données (Cf. Daudin et *al.* [9, 8]) selon les critères définis au paragraphe précédent.

- Dans le premier jeu de données, 8 variables caractérisent la composition de 85 conteneurs de lait.

- Dans le second jeu de données, la prise de poids hebdomadaire de 28 chèvres est mesurée durant 11 semaines, la première variable étant le poids noté la première semaine.

Pour ces deux exemples, 100 sous-échantillons ont été tirés, par la méthode du Bootstrap, à partir de l'échantillon d'origine.

a Composition du lait

Résultats :

L'ACP du premier jeu de données se distingue par un premier axe expliquant 72 % de la variabilité (Cf. Figure 1). Les critères de stabilité mesurés pour chacune des méthodes de Bootstrap $[A]$ et $[B]$, présentées dans le paragraphe 5.4, sont indiquées dans le tableau 1. Enfin, l'évolution des $A_k^{(b)}$ (avec $b = 1, \ldots, 100$ et $k = 1, \ldots, 8$) est représentée sur la figure 2 pour la méthode $[A]$ et sur la figure 3 pour la méthode $[B]$.

Interprétation :

Il faut tout d'abord souligner que la stabilité parfaite du sous-espace de dimension p est naturelle car E_p est engendré par l'ensemble des variables.

Il semble que l'on puisse tenir compte des quatre premiers axes sans risquer qu'un nouvel échantillonnage sur la même population n'aboutisse à des résultats différents.

On doit cependant noter que les valeurs propres associées aux deux ou trois premiers axes ont une variance élevée, ce qui tendrait à prouver que si leur direction est stable, leur part d'explication de la variabilité globale est sujette à caution.

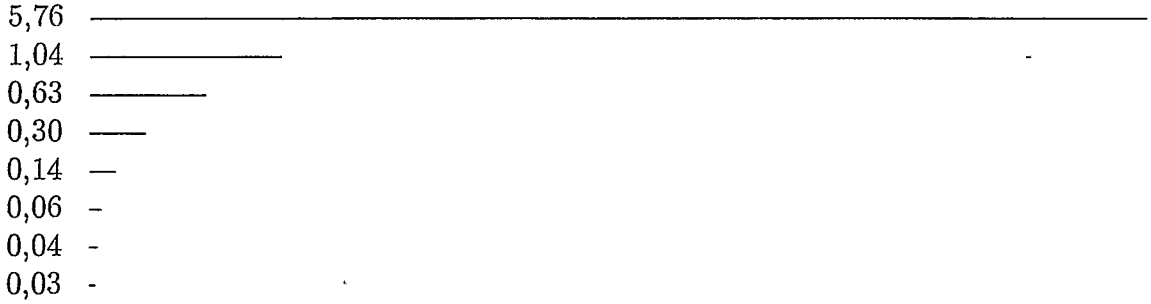


Figure 1 : Valeurs propres de l'ACP sur l'échantillon des données de composition du lait.

Dimension	Méthode [A]						Méthode [B]				
k	λ_k	$S_{\lambda k}$	Biais	A_k	S_{Ak}	EQM$_k$	λ_k	$S_{\lambda k}$	A_k	S_{Ak}	EQM$_k$
1	5,76	0,29	0,00	1,00	0,00	0,00	5,74	1,19	1,00	0,00	0,00
2	1,07	0,14	0,03	1,96	0,05	0,06	1,02	0,12	1,96	0,07	0,08
3	0,63	0,12	0,00	2,99	0,02	0,02	0,60	0,09	2,98	0,03	0,04
4	0,30	0,06	0,00	3,99	0,03	0,03	0,28	0,04	3,98	0,03	0,04
5	0,14	0,04	0,00	4,96	0,12	0,13	0,14	0,03	4,99	0,03	0,03
6	0,06	0,02	0,00	5,76	0,29	0,38	0,06	0,02	5,76	0,03	0,24
7	0,03	0,01	-0,01	6,74	0,24	0,35	0,03	0,01	6,77	0,02	0,23
8	0,02	0,01	-0,01	8,00	0,00	0,00	0,02	0,01	8,00	0,00	0,00

Tableau 1 : Critères de stabilité des valeurs propres et des sous-espaces de représentation mesurés sur les données de composition du lait (calculés pour 100 sous-échantillons).

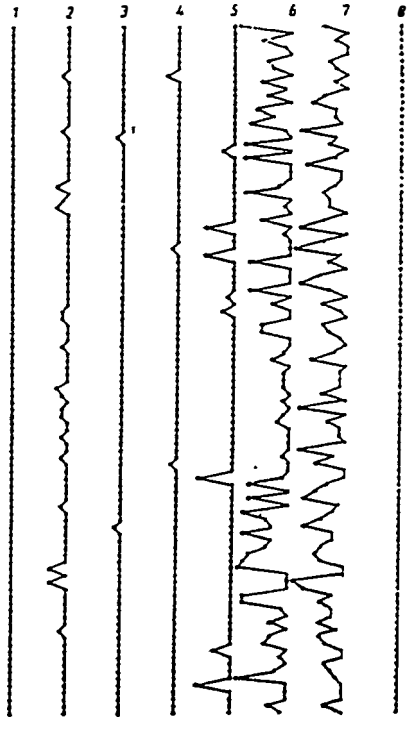


Fig. 2 : Représentation de $A_k^{(b)}$ (Méthode [A])

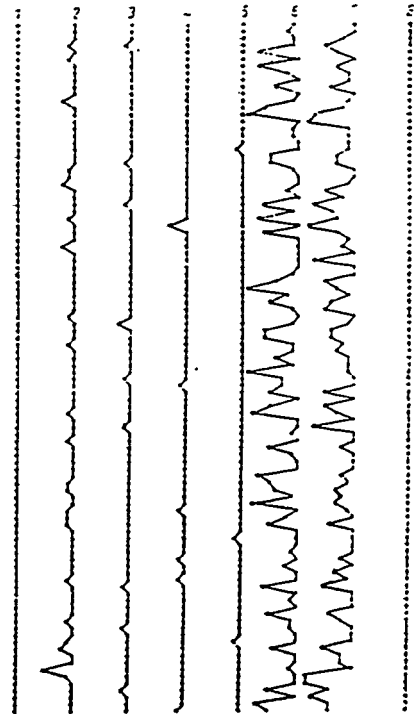


Fig. 3 : Représentation de $A_k^{(b)}$ (Méthode [B])

b Accroissements du poids des chèvres

Résultats :

Les valeurs propres de l'ACP du second jeu de données sont indiquées sur la figure 4. Les critères de stabilité mesurés pour les deux méthodes de Bootstrap [A] et [B] sont présentés dans le tableau 2. Comme les résultats obtenus pour ces deux méthodes sont similaires, seule l'évolution des $A_k^{(b)}$ pour la méthode [B] est visualisée (figure 5).

Interprétation :

L'interprétation des résultats des simulations réalisées à partir du second lot de données est malheureusement sans appel. Le seul sous-espace relativement stable est le sous-espace de dimension 10 ($\mathrm{EQM}_{10} = 0,05$).

Ceci conduit l'utilisateur à n'apporter aucun crédit à toute représentation graphique de ces données. On peut, par exemple, imaginer que l'expérimentation a été largement perturbée par la présence d'individus malades.

Dans un tel contexte, la seule utilité de l'ACP se limite à la détection de données aberrantes. Toute conclusion concernant les variables serait hasardeuse.

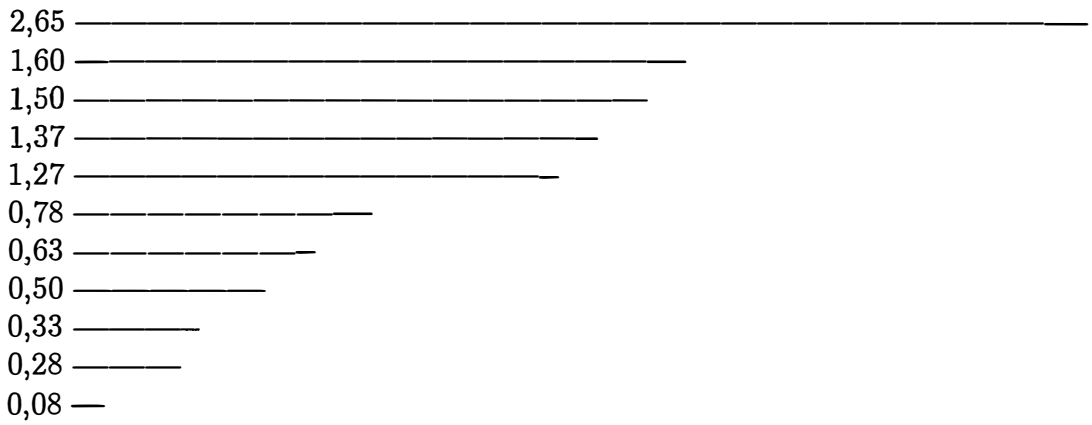


Figure 4 : Valeurs propres de l'ACP sur l'échantillon des données d'accroissement du poids des chèvres.

Dimension	Méthode [A]						Méthode [B]				
k	λ_k	$S_{\lambda k}$	Biais	A_k	S_{Ak}	EQM_k	λ_k	$S_{\lambda k}$	A_k	S_{Ak}	EQM_k
1	3,02	0,37	0,37	0,78	0,25	0,33	3,06	0,66	0,77	0,23	0,32
2	2,12	0,25	0,52	1,35	0,33	0,73	2,00	0,32	1,35	0,31	0,72
3	1,64	0,16	0,14	2,21	0,30	0,84	1,50	0,22	2,22	0,28	0,83
4	1,29	0,15	-0,08	3,33	0,26	0,72	1,15	0,19	3,31	0,25	0,73
5	0,99	0,14	-0,28	4,61	0,22	0,45	0,85	0,16	4,56	0,21	0,49
6	0,71	0,12	-0,07	5,53	0,23	0,52	0,60	0,12	5,55	0,20	0,49
7	0,50	0,10	-0,13	6,61	0,19	0,43	0,42	0,08	6,60	0,15	0,43
8	0,34	0,08	-0,16	7,65	0,19	0,40	0,29	0,08	7,69	0,17	0,35
9	0,22	0,06	-0,11	8,64	0,19	0,41	0,18	0,05	8,66	0,16	0,38
10	0,12	0,05	-0,16	9,97	0,04	0,05	0,11	0,04	9,97	0,04	0,05
11	0,04	0,02	-0,04	11,00	0,00	0,00	0,03	0,02	11,00	0,00	0,00

Tableau 2 : Critères de stabilité des valeurs propres et des sous-espaces de représentation mesurés sur les données d'accroissement du poids des chèvres (calculés pour

100 sous-échantillons).

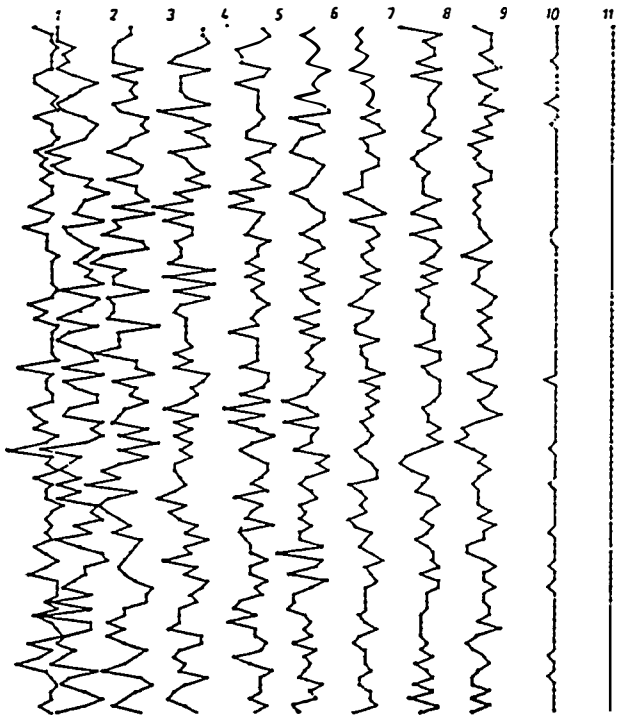


Fig. 5 : Représentation de $A_k^{(b)}$ (Méthode $[B]$)

c Conclusions générales

• La méthode $[B]$ surestime la variance de λ_1 par rapport à la méthode $[A]$. La différence entre les deux méthodes est particulièrement flagrante dans le premier exemple. De plus, pour cet exemple, l'estimation de EQM_k obtenue avec la méthode $[B]$ est supérieure à celle obtenue avec la méthode $[A]$ lorsque $k = 1, 2$ et 3.

Ces observations sont dues à l'approximation de la matrice des corrélations R_X par la matrice de variance-covariance V_W, consentie dans l'utilisation de la méthode $[B]$, basée sur le Bootstrap appliqué aux variables de l'ACP.

• On constate que $\bar{\lambda}_1^{(b)} \geq \lambda_1$ pour la méthode $[A]$. Ceci est en parfait accord avec les considérations théoriques exposées dans le chapitre 3. A l'inverse, le biais mesuré pour la dernière valeur propre est négatif $\bar{\lambda}_p^{(b)} \leq \lambda_p$.

• La règle empirique de détermination de la dimension du sous-espace de représentation en fonction de l'effondrement soudain des valeurs propres est d'utilisation courante. Elle consiste à choisir le sous-espace de représentation k lorsqu'un écart important entre λ_k et λ_{k+1} est observé. Cette règle simple n'est pas vérifiée par les deux exemples étudiés.

Dans le premier exemple, elle conduirait à retenir les sous-espaces de dimension 1 ou

3. En effet, la proximité des valeurs de λ_2 et λ_3 conduirait à rejeter le sous-espace de dimension 2. La même considération est vraie pour λ_4 et λ_5. Or les estimations de la stabilité par la méthode du Bootstrap, telles que EQM, montrent que les quatre premiers axes sont stables.

Dans le second exemple, la même règle conduirait à choisir les sous-espaces de dimension 1 ou 5. Or ces deux sous-espaces sont très instables.

Le critère de choix du nombre de composantes basé sur "l'éboulis" des valeurs propres (*screegraph*) ne peut donc, en aucun cas, se substituer à l'étude nécessaire de la stabilité d'une ACP.

• Enfin, la stabilité des derniers axes de l'ACP est souvent discutée. Dans le premier exemple, les premières variables de l'ACP sont stables tandis que les dernières sont instables. Au contraire, dans le second exemple, seules les dernières variables de l'ACP sont stables.

A l'instar du critère précédent, aucune règle basée sur le numéro des axes ne peut se substituer à l'étude de la stabilité d'une ACP, compte tenu des résultats obtenus.

La validation croisée

6.1 Présentation de la validation croisée

Comme le Jacknife et le Bootstrap, la validation croisée est une méthode de rééchantillonnage. Cependant, le but poursuivi est différent.

a Définition

Soit un modèle construit et ajusté à partir de données. La validation croisée consiste à évaluer les qualités de prédiction du modèle sur de nouvelles données indépendantes de celles qui ont servi à le construire.

Le modèle sera déclaré valide si ses qualités sont suffisantes pour ce qu'on veut en faire. La validation ne constitue pas une épreuve de vérité.

b Justification intuitive

Un modèle sert à prédire une variable, habituellement à l'aide de prédicteurs. C'est le cas en régression, ainsi qu'en analyse discriminante. C'est aussi vrai pour l'interpolation d'une surface, où l'on cherche à estimer la valeur d'une variable Z en un point, à partir de la position de ce point définie par ses coordonnées $t = (x, y)$.

Pour construire un modèle, nous avons à notre disposition un ensemble de couples (variable Z, prédicteurs t). Pour que ce modèle ait quelque utilité, il nous faut supposer que ces couples sont tirés au hasard dans une population, et qu'à l'avenir, ceux qui se présenteront seront tirés au hasard de même. Les données constituent un échantillon.

La modélisation va consister à déduire de cet échantillon une fonction $\hat{z} = f(t)$, qui sera le prédicteur de Z pour toute nouvelle valeur de t. Le prédicteur sera jugé bon si les valeurs prédites z s'écartent peu des vraies valeurs z, les écarts étant appelés erreurs de prédiction. Certaines erreurs étant plus grandes que d'autres, on juge un modèle sur l'ensemble des erreurs qu'il peut commettre sur son domaine d'application. Les critères les plus courants sont déduits des deux premiers moments de ces erreurs :

Biais	$\mu_1 = E\left(Z - \hat{z}\right)$
Erreur quadratique moyenne	$\mu_2' = E\left(Z - \hat{z}\right)$
Variance	$\mu_2 = \mu_2' - \mu_1^2$
$E(.)$ désigne l'espérance	

Pour évaluer ces critères, on peut utiliser les écarts entre les valeurs z et leurs prédictions $\hat{z}$ sur l'échantillon dont on dispose. Toutefois, les écarts d'un modèle aux données qui ont servi à l'ajuster sont en général plus faibles que les écarts à de nouvelles données qui n'auraient pas participé à l'ajustement. Les premiers sont appelés erreurs apparentes, les seconds erreurs vraies. Ce sont les seconds qui nous intéressent.

Le principe de la validation croisée est de séparer au hasard les données servant à la construction et à l'ajustement du modèle (phase appelée apprentissage), de celles qui servent à l'évaluation des erreurs (figure 1). On place ainsi le modèle dans ses conditions supposées d'utilisation : construction à partir d'un échantillon tiré au hasard dans une population, puis utilisation sur des données issues de la même population.

c Domaines d'utilisation

Les méthodes linéaires comme la régression et l'interpolation par krigeage permettent d'évaluer directement les moments des erreurs de prédiction, pour autant que le modèle soit vrai, et sa forme connue *a priori*, sans avoir connaissance des données.

Au contraire, les méthodes où l'on choisit régresseurs ou transformations de variables d'après les données, où l'on élimine des points aberrants, ..., ne permettent pas ce calcul direct. La régression non-linéaire, la segmentation, ne le permettent pas non plus.

La validation croisée est un moyen d'estimer les moments des erreurs de prédiction quand le modèle utilisé ne permet pas de le faire directement, ou quand les hypothèses que l'on doit admettre pour cela sont à mettre en doute.

d Variantes de la validation croisée

Procédé original

On tire au hasard la moitié des données pour l'apprentissage, et l'on réserve l'autre moitié à la validation. Le premier échantillon est appelé échantillon d'apprentissage, le second échantillon-test.

Pour la validation, le modèle n'est donc ajusté que sur la moitié des données. Ensuite,

si le modèle donne satisfaction, on l'ajuste définitivement sur l'ensemble des données, pour ne pas se priver inutilement de l'échantillon-test.

Cette procédure donne donc une idée pessimiste des performances du modèle final. Pour se rapprocher de celles-ci, on peut réserver moins de 50 % des données à la validation, mais alors l'évaluation des moments des erreurs sera moins précise.

Variante

Plutôt que d'écarter une partie non négligeable des données, on peut éliminer chaque individu à tour de rôle, et le prédire à l'aide de tous les autres individus. On approche ainsi beaucoup mieux les performances réelles du modèle. Mais il faut répéter autant de fois l'apprentissage complet qu'il y a d'individus.

Si cet apprentissage est automatique, ce n'est qu'une question de temps de calcul. Par exemple, pour valider une régression multiple pas-à-pas faite sur n individus, on répétera la procédure de sélection de variables sur chacun des n échantillons possibles de $n-1$ individus.

En revanche, si l'apprentissage nécessite l'intervention de l'homme, il n'est pas possible de le répéter n fois, en oubliant à chaque fois ce qui vient d'être fait. Pour cette raison, cette variante, préférable au procédé original, n'est pas applicable dans tous les cas.

6.2 Application à l'interpolation par Krigeage

a Présentation sommaire du krigeage

Le krigeage est un procédé d'interpolation dans l'espace de mesures faites en un nombre fini de points.

Pour être krigeables, les données doivent varier de manière suffisamment lisse. Formellement, l'espérance du carré de l'écart entre deux mesures faites en des points différents ne doit être fonction que de la distance qui sépare ces points, et éventuellement de l'orientation du segment qui les joint. C'est l'hypothèse dite intrinsèque.

Habituellement, cette fonction est croissante, car des mesures faites en des points proches sont en général plus semblables que celles faites en des points éloignés.

La connaissance de cette fonction, appelée variogramme, permet de déduire un interpolateur linéaire optimal. La variable Z est alors estimée en tout point (x_0, y_0) par une moyenne pondérée $\hat{z}_0$.

Les coefficients de cette pondération sont fonction des positions des mesures dispo-

nibles $[(x_i, y_i), i = 1 \ldots, n]$, et du nouveau point (x_0, y_0).

L'estimateur $\hat{z}_0$ est sans biais, et de variance minimale connue. On peut donc calculer en tout point l'erreur-type, ou écart- type de l'erreur de krigeage, écart entre la valeur z_0 (inconnue) et sa prédiction $\hat{z}_0$. De même que pour la régression, cet écart-type dépend des prédicteurs x_0 et y_0 : la prédiction sera d'autant plus précise qu'il y a des points mesurés au voisinage du point à prédire. Cette erreur-type n'est exacte que si le modèle statistique des variations de Z dans l'espace est vrai : hypothèse intrinsèque vraie, et variogramme connu exactement.

En pratique, on ne connaît pas le variogramme vrai, mais on l'estime à partir des données. Pour cela, on trace des variogrammes empiriques, qui sont des écarts quadratiques moyens observés par classe d'orientation et de distance, puis on ajuste une fonction au jugé. En effet, les méthodes statistiques d'ajustement de fonctions sont trop peu robustes pour les variogrammes, sensibles aux valeurs aberrantes. De plus, un bon ajustement doit être privilégié pour les faibles distances, car les poids des voisins les plus proches sont prépondérants dans les interpolations. L'ajustement d'un variogramme suppose une expertise pas suffisamment formalisée pour être automatisable.

Les données servent donc à la fois pour l'estimation du variogramme (manuelle), et l'interpolation proprement dite (automatique).

b Objet et mise en oeuvre de la validation croisée

La validation croisée va permettre ici de vérifier si l'interpolation est bien sans biais, et si les erreurs-types annoncées sont bien celles qu'on peut attendre en réalité. Si c'est le cas, on acceptera comme valides l'hypothèse intrinsèque et le variogramme ajusté manuellement.

Le procédé de validation croisée communément admis pour le krigeage, et programmé dans certains logiciels (Englund, Sparks, 1988) [5], est décrit par Davis (1987) [1] :

> *Select a model and associated parameters for the underlying variogram. Leave one sample point out of the data set and estimate the value at that location by kriging with the selected variogram model and the remaining data. Compute the difference between the actual and estimated value for each data location. Divide each difference squared by the kriging variance of the estimate. Compute the mean and standard deviation of these difference data. Repeat the above for each model or parameter set candidate. Choose that model or set of model parameters that causes the mean of the difference data to be closest to zero and the standard deviation closest to one.*

Le critère de validité est donc que l'erreur réduite (c'est-à-dire divisée par l'écart-type d'estimation) doit être de moyenne proche de 0, et d'écart-type proche de 1.

D'après ce que nous avons dit plus haut, cette seconde variante de la validation croisée n'est pas applicable dans le cas du krigeage, puisque l'estimation du variogramme est manuelle. Elle est néanmoins conduite en ne remettant pas en cause le variogramme, ajusté une fois pour toutes sur l'ensemble des données (figure 2).

Cette variante de la validation croisée n'est théoriquement pas adaptée au procédé d'interpolation dans son ensemble. Nous avons voulu savoir si cela impliquait des erreurs considérables sur deux exemples où nous disposons d'une grille complète de points mesurés, que nous nous proposons de retrouver par interpolation à partir d'un échantillon tiré au hasard. Les données ont été aimablement communiquées par Monsieur Guillobez (CIRAD-CA). Il s'agit de mesures de rugosité du sol par relevé d'altitude au laser à tous les noeuds d'une grille 30x30, de maille 12x12 mm. L'appareil de mesure a été mis au point au CIRAD-SAR.

c Résultats

Le premier jeu de données (try1) est relevé sur un sol peu rugueux (figure 3). L'échantillon tiré au hasard comporte 178 points (figure 4). Sur cet échantillon, le variogramme ajusté est linéaire, et n'a pas la même portée dans le sens Est-Ouest (figure 5) que dans le sens Nord-Sud (figure 6). La surface interpolée (figure 7) a la même forme générale que la surface vraie, tout en étant nettement plus lisse. La différence constitue l'erreur de krigeage.

La validation croisée telle que présentée plus haut (méthode présentée par Davis) nous invite à être confiant : les erreurs réduites ont un écart-type inférieur à 1, et une moyenne quasiment nulle.

Dans une deuxième étape, revenons maintenant au procédé original de validation croisée, et regardons la distribution des erreurs réduites sur l'ensemble des 30x30 - 172 = 728 données qui n'ont pas servi au calcul du variogramme, ni à l'interpolation. Remarquons la dissymétrie de l'histogramme des erreurs réduites (figure 9). L'écart-type obtenu est légèrement plus grand que précédemment.

Le second jeu de données est relevé sur un sol motteux et caillouteux, nettement plus rugueux que le premier (figure 10). On reprend le même échantillon de points que précédemment, ils sont donc disposés comme indiqué figure 4. Les variogrammes ajustés suivant les sens Nord-Sud et Est-Ouest montrent une structure moins accusée que dans le premier cas, sans être toutefois erratique (figures 11 et 12).

La validation croisée suivant la méthode présentée par Davis montre des erreurs réduites assez bien distribuées, de moyenne nulle, et d'écart-type acceptable : 1,167.

La deuxième étape, de comparaison aux 728 données hors échantillon, montre cette fois-ci des erreurs pas tout à fait bien centrées, dont l'écart-type, 1,51, dépasse nettement l'estimation par la méthode usuelle (figure 15).

6.3 Conclusion

L'absence de prise en compte de la variabilité de la fonction estimant le variogramme dans le procédé habituel de validation conduit sur ces deux exemples à être trop confiant dans la méthode du krigeage. Le procédé original de validation croisée est au contraire conservatif. La mise au point de méthodes automatiques d'ajustement du variogramme permettrait d'en appliquer convenablement la variante, et d'avoir une idée plus réaliste de la précision des interpolations.

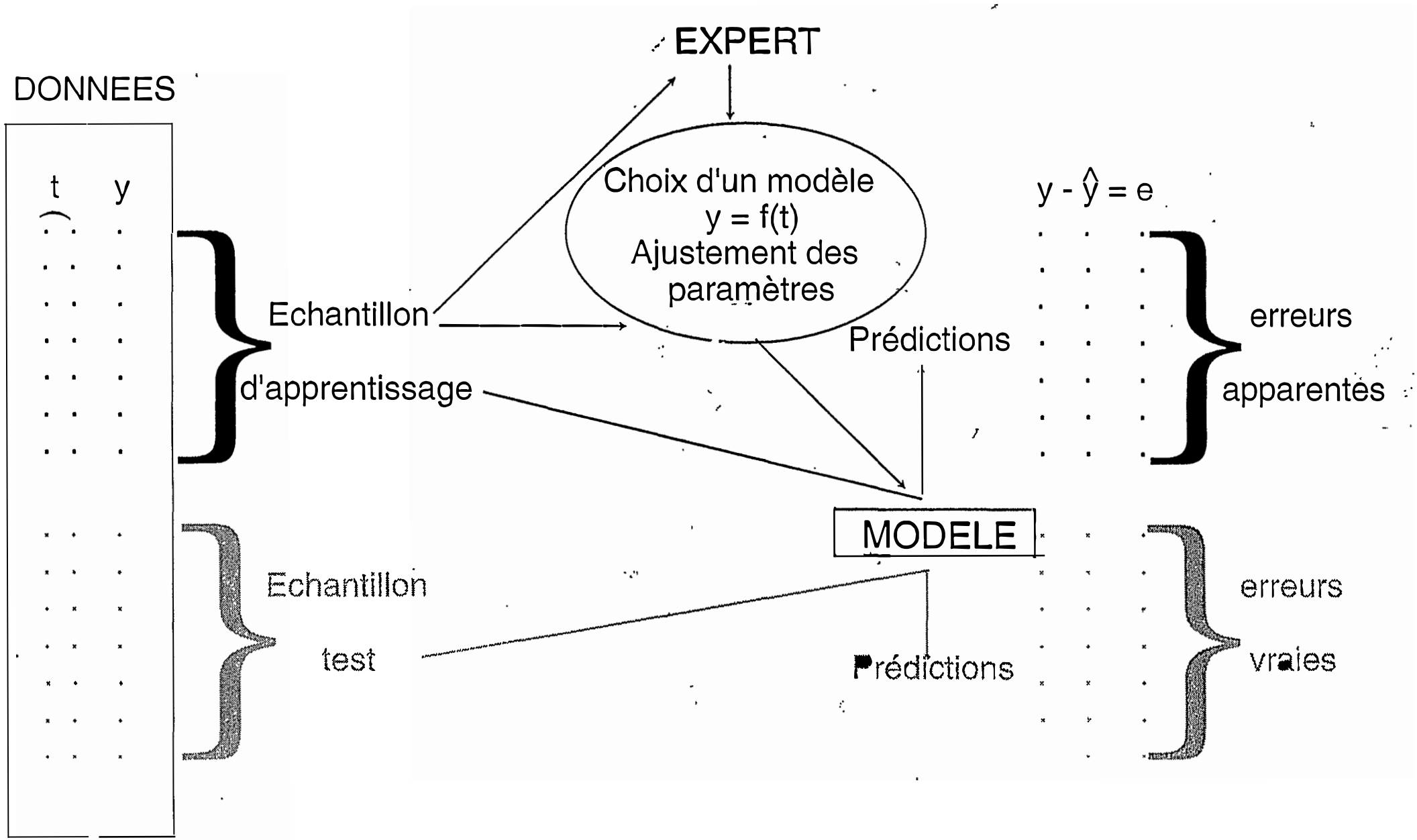


Figure 1 : Principe de la validation croisée

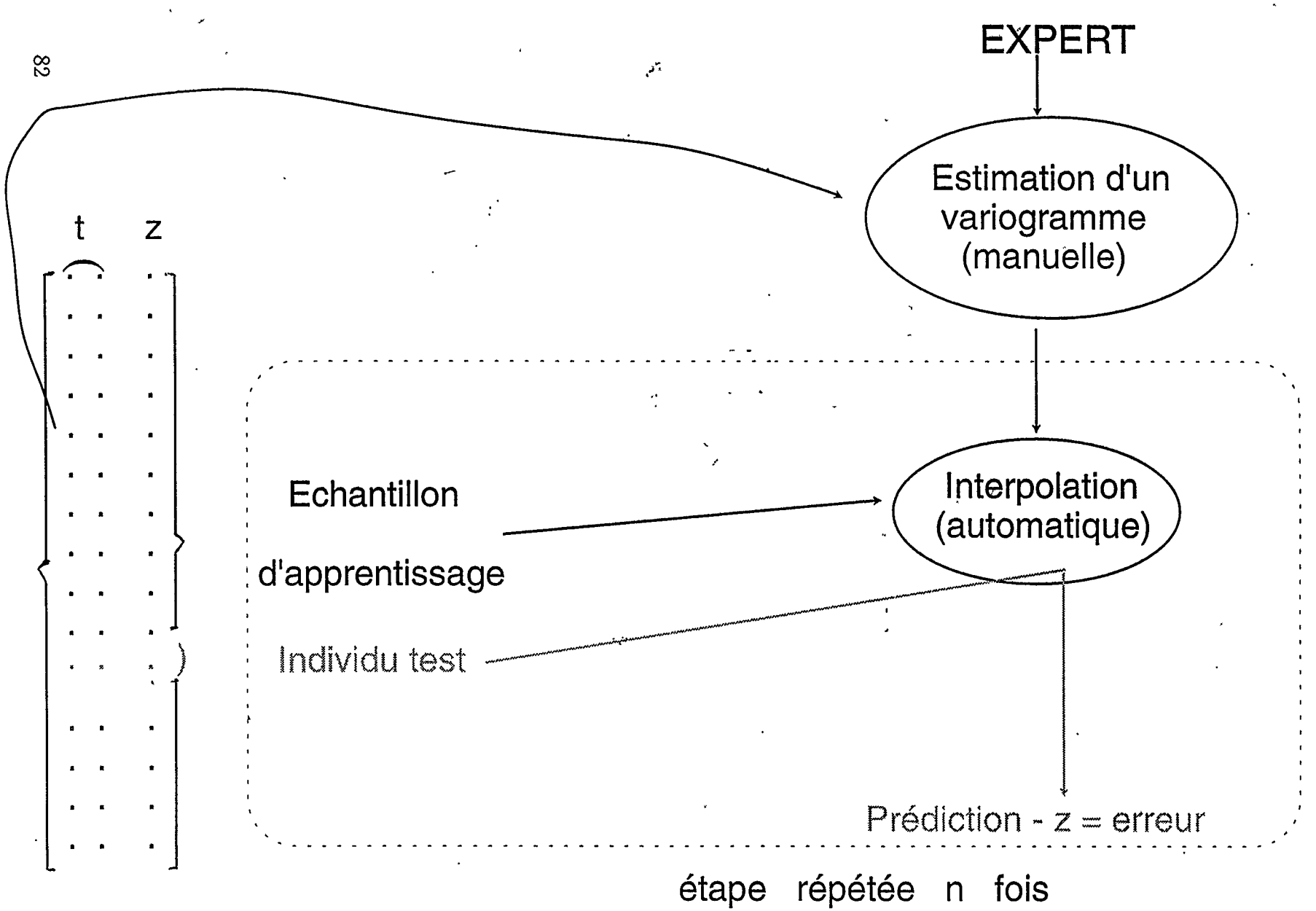


Figure 2 : Principe de la validation croisée habituellement pratiquée par le krigeage

BLOC – DIAGRAMME DES DONNEES NON INTERPOLEES

DONNEES DE try1

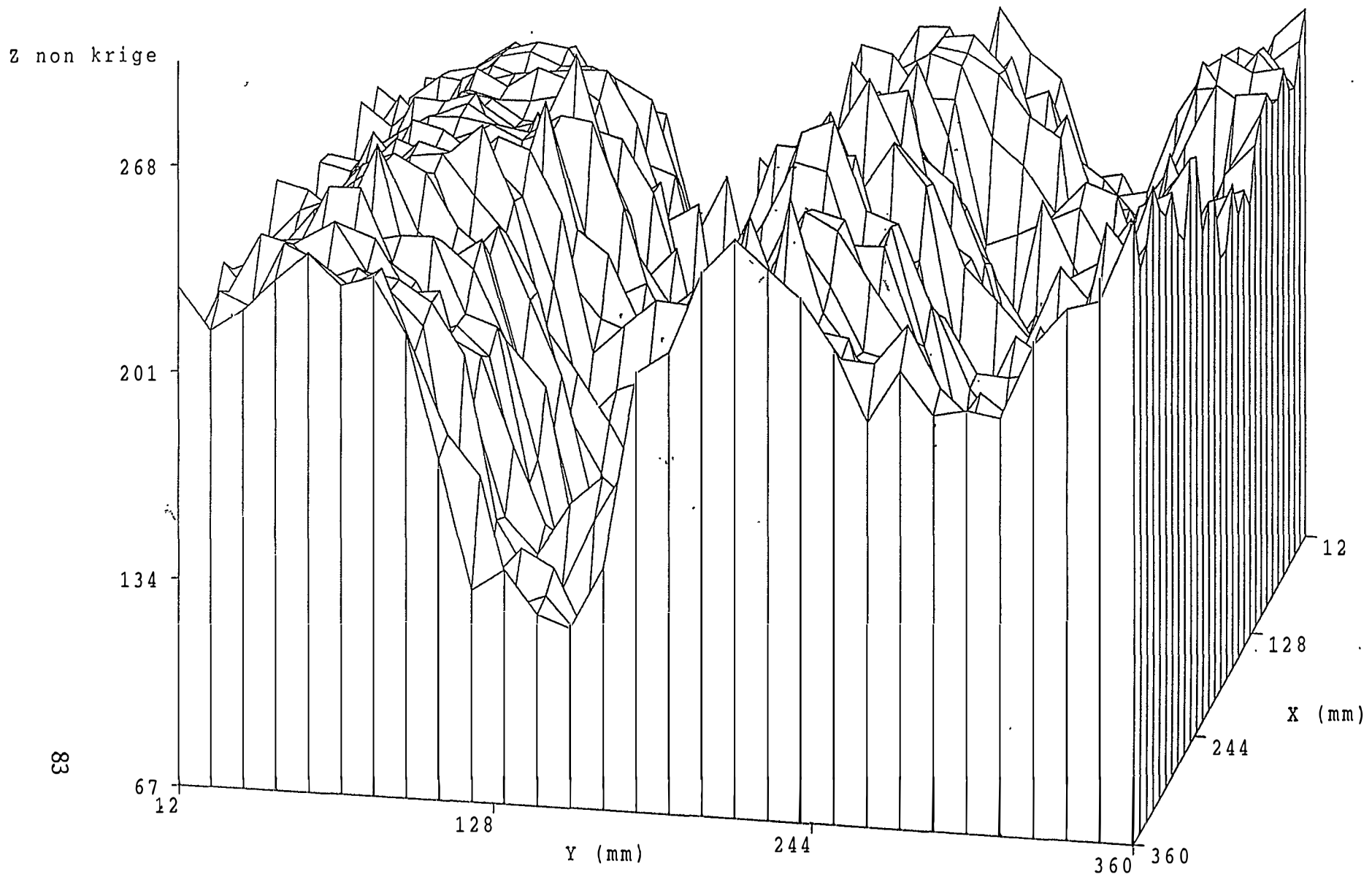
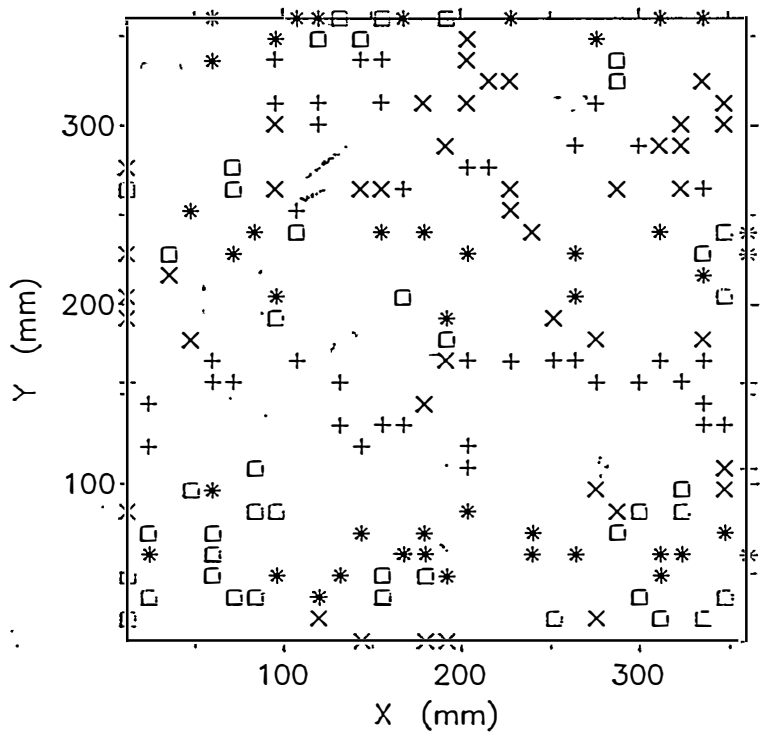


FIGURE 3

Figure 4



1st Quartile:	82.440 ≦ + ≦ 179.930
2nd Quartile:	179.930 < × ≦ 208.560
3rd Quartile:	208.560 < □ ≦ 226.300
4th Quartile:	226.300 < ✳ ≦ 257.270

FIGURE 5 : Variogramme Est-Ouest

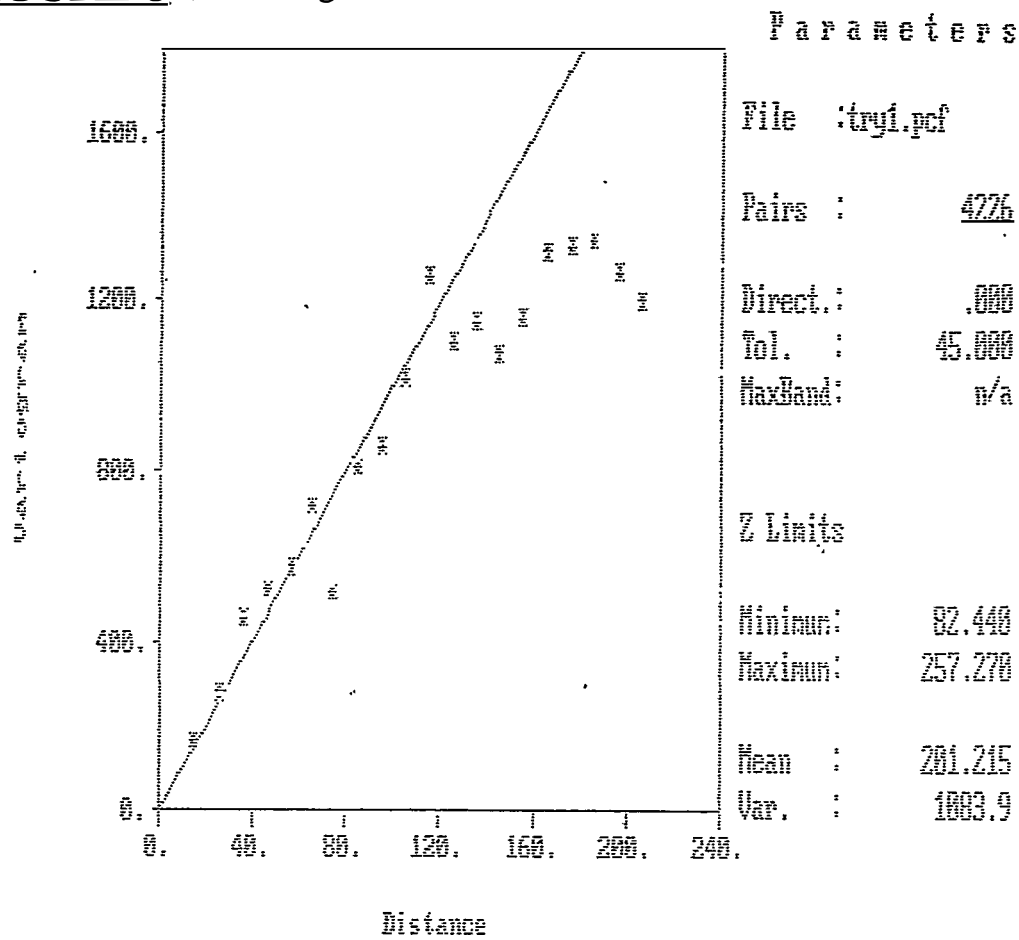
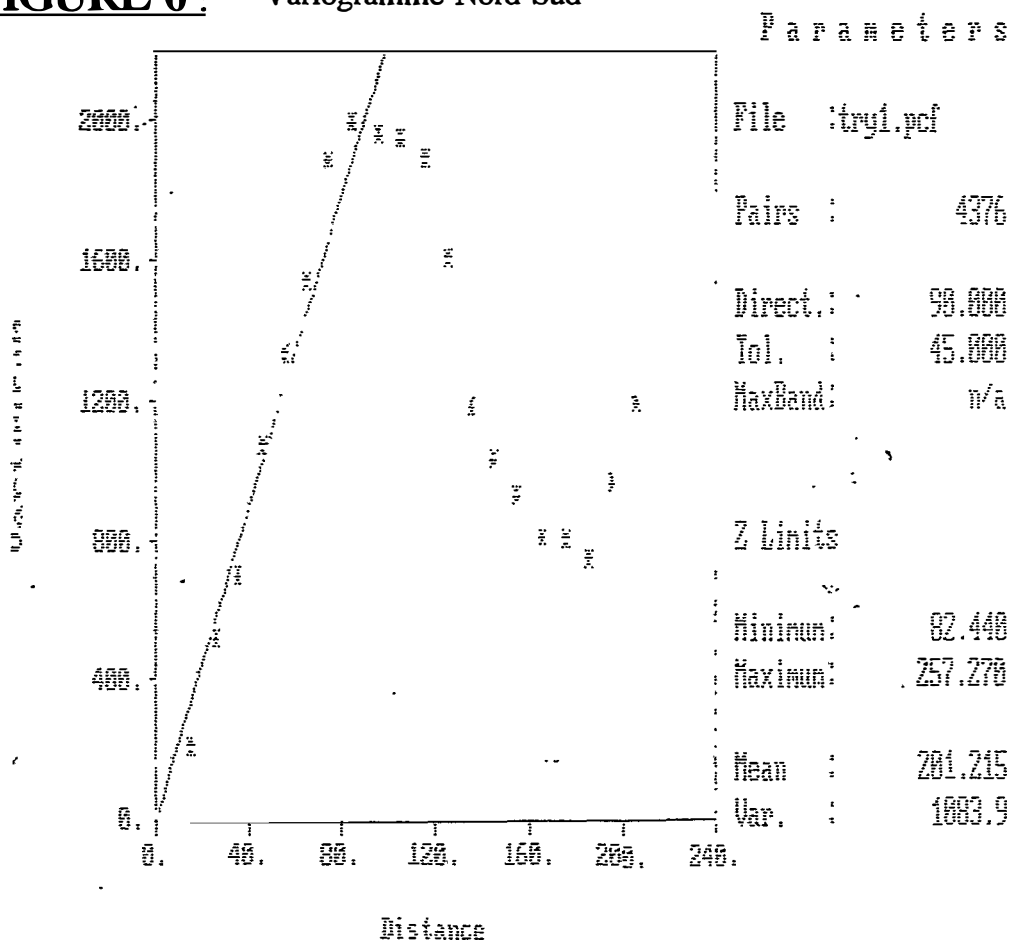


FIGURE 6 : Variogramme Nord-Sud



BLOC DIAGRAMME DES DONNEES KRIGEES

DONNEES DE try1

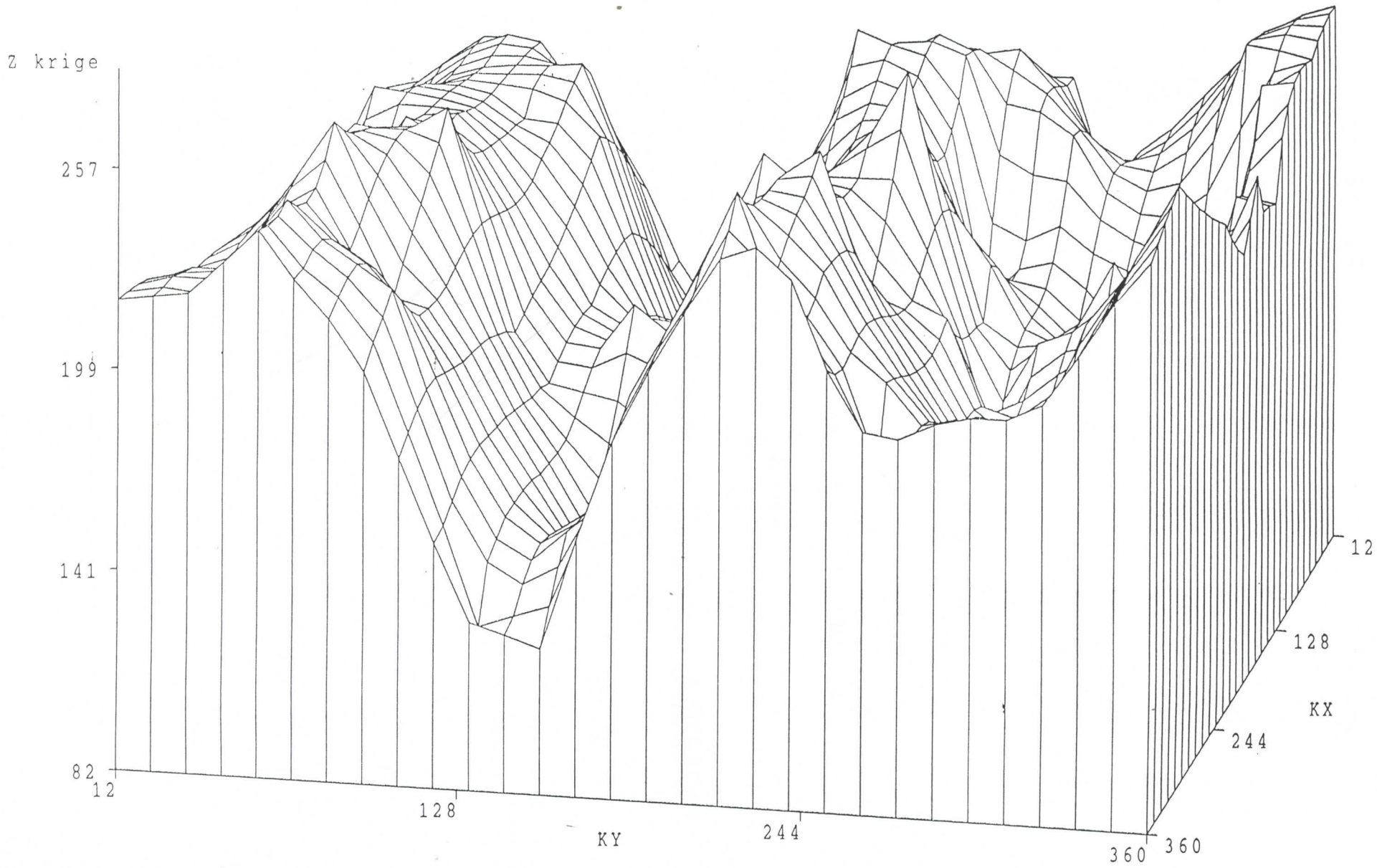


FIGURE 7

FIGURE 8 : Description de la distribution des erreurs réduites sur les données de l'échantillon

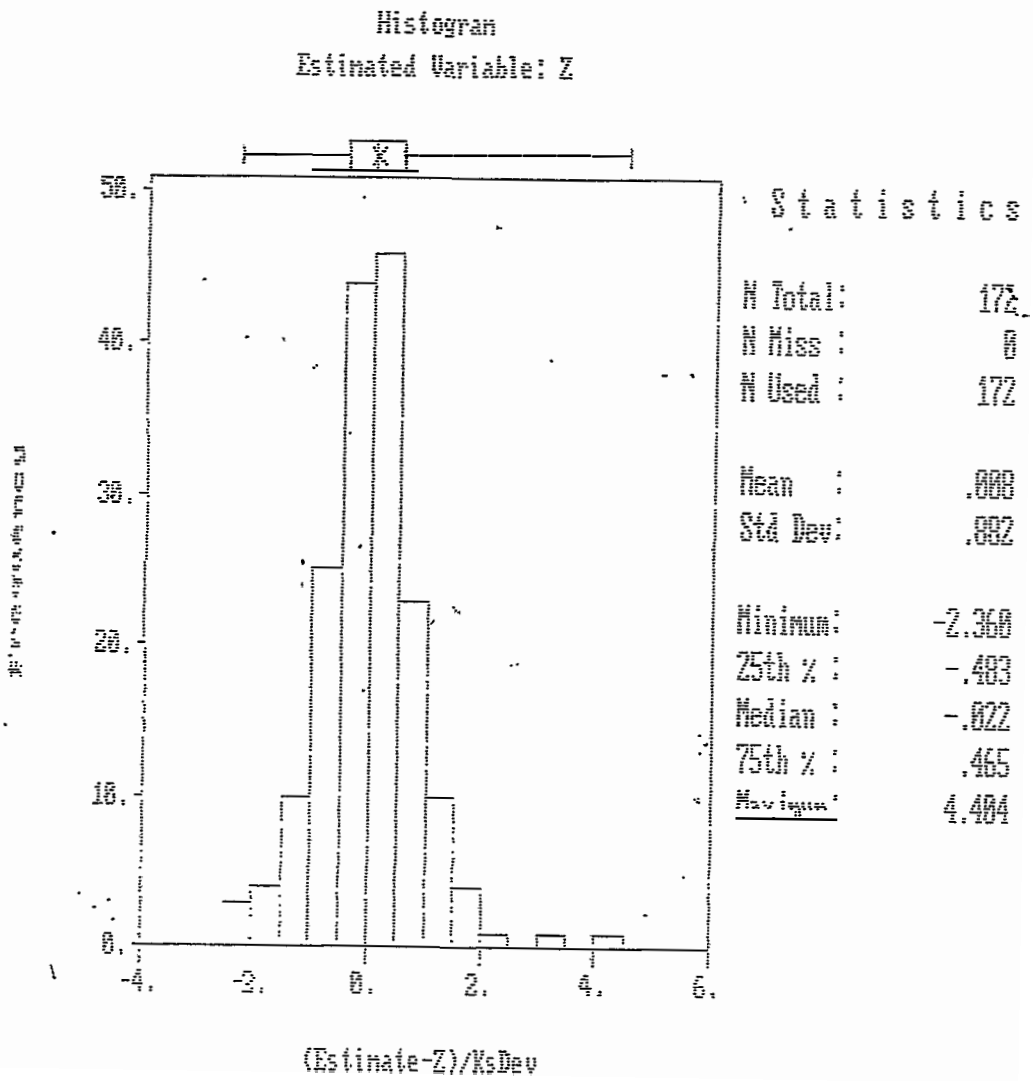


FIGURE 9 : Description de la distribution des erreurs réduites sur les individus hors échantillon

VALIDATION CROISEE : DEUXIEME PHASE 11:42 Sunday, September 5, 1993 4
COMPARAISON DE TOUTE LA GRILLE KRIGEE AVEC LA GRILLE VRAIE
DONNEES DE try1

Univariate Procedure

Variable=SERRZ Erreur de krigeage reduite

Moments

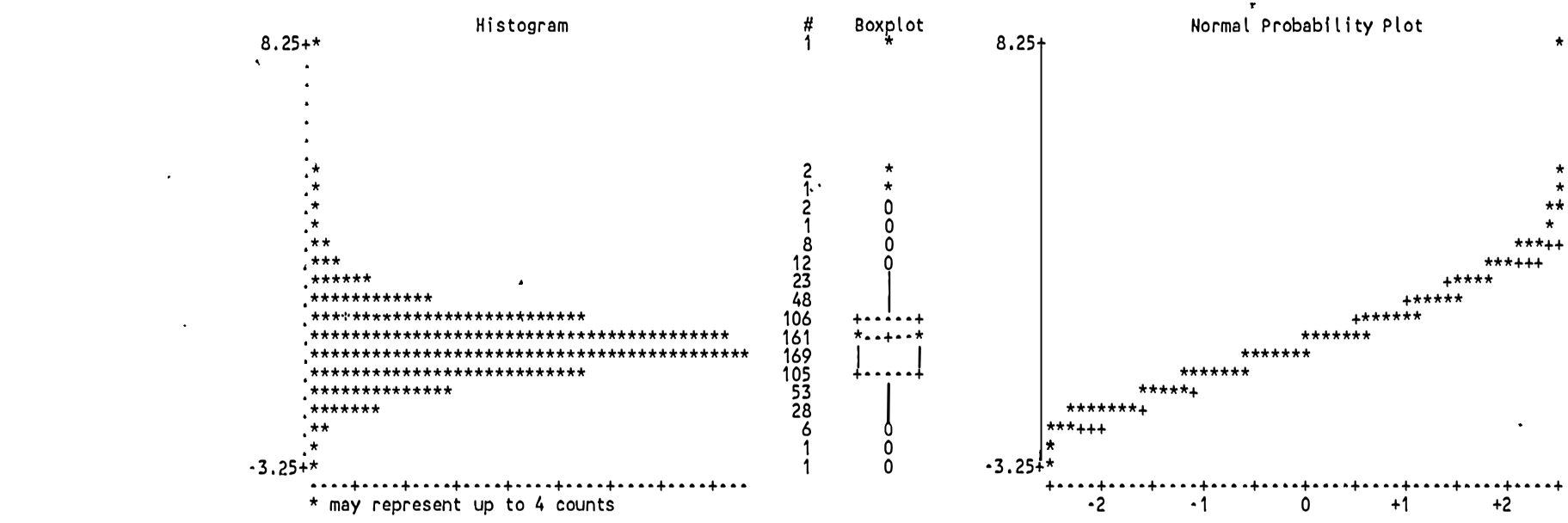
N	728	Sum Wgts	728
Mean	0.066558	Sum	48.45416
Std Dev	1.036799	Variance	1.074953
Skewness	1.303211	Kurtosis	7.290923
USS	784.7157	CSS	781.4907
CV	1557.74	Std Mean	0.038426
T:Mean=0	1.732091	Prob>|T|	0.0837
Num ^= 0	728	Num > 0	365
M(Sign)	1	Prob>|M|	0.9704
Sgn Rank	3453	Prob>|S|	0.5433
W:Normal	0.948751	Prob<W	0.0001

Quantiles(Def=5)

100% Max	8.448094	99%	2.978565
75% Q3	0.568066	95%	1.679772
50% Med	0.005105	90%	1.222579
25% Q1	-0.53635	10%	-1.13283
0% Min	-3.08767	5%	-1.46318
		1%	-2.0086
Range	11.53577		
Q3-Q1	1.104415		
Mode	-3.08767		

Extremes

Lowest	Obs	Highest	Obs
-3.08767(	268)	3.964407(	451)
-2.63784(	462)	4.329628(	423)
-2.31358(	383)	4.690922(	142)
-2.23415(	8)	4.908066(	424)
-2.08(	379)	8.448094(	399)



BLOC – DIAGRAMME DES DONNEES NON INTERPOLEES

DONNEES DE jp1

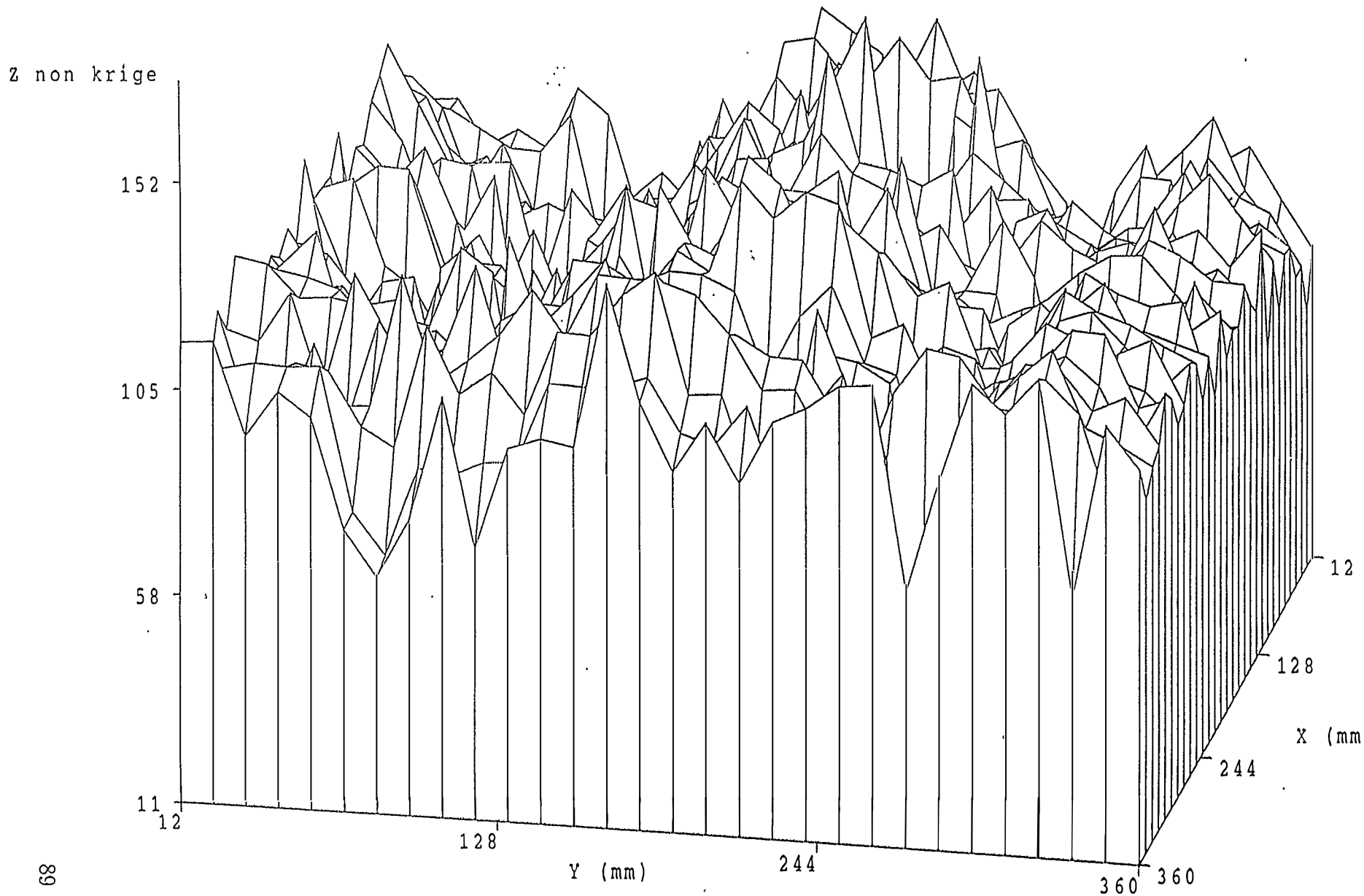


FIGURE 10

FIGURE 11 : Variogramme Est-Ouest

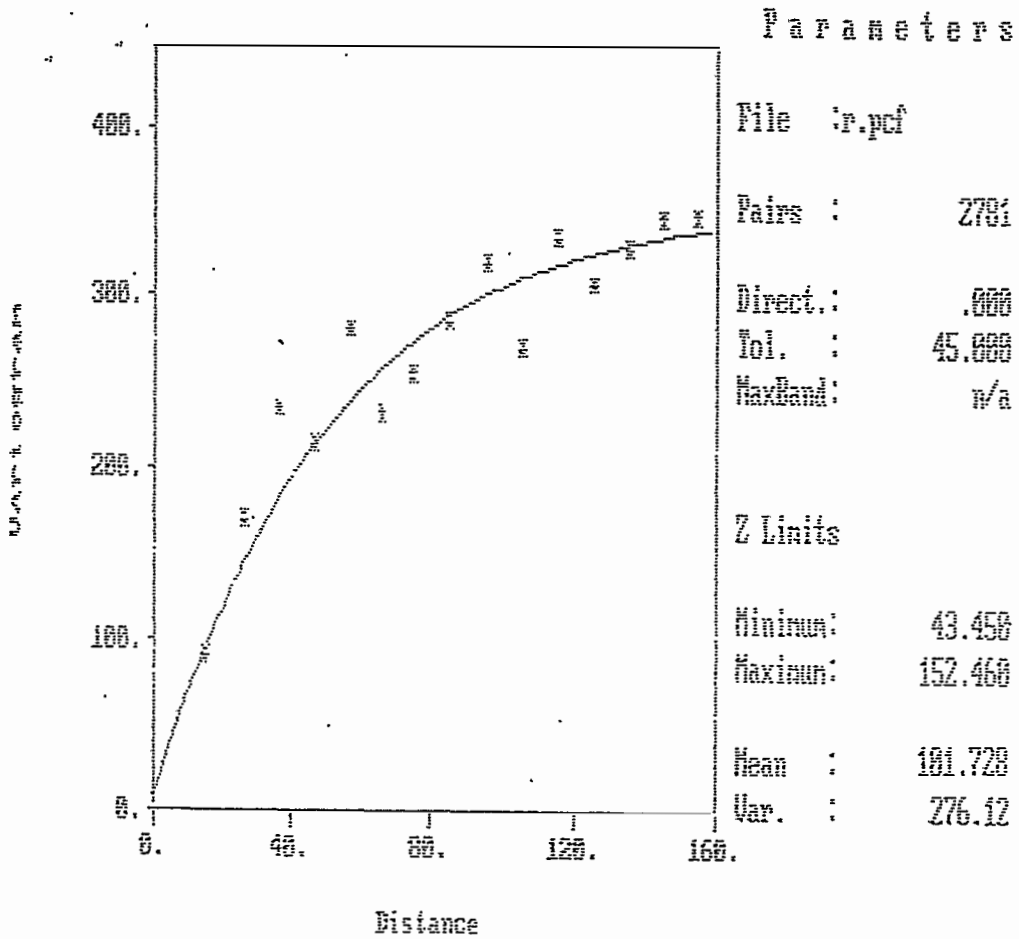


FIGURE 12 : Variogramme Nord-Sud

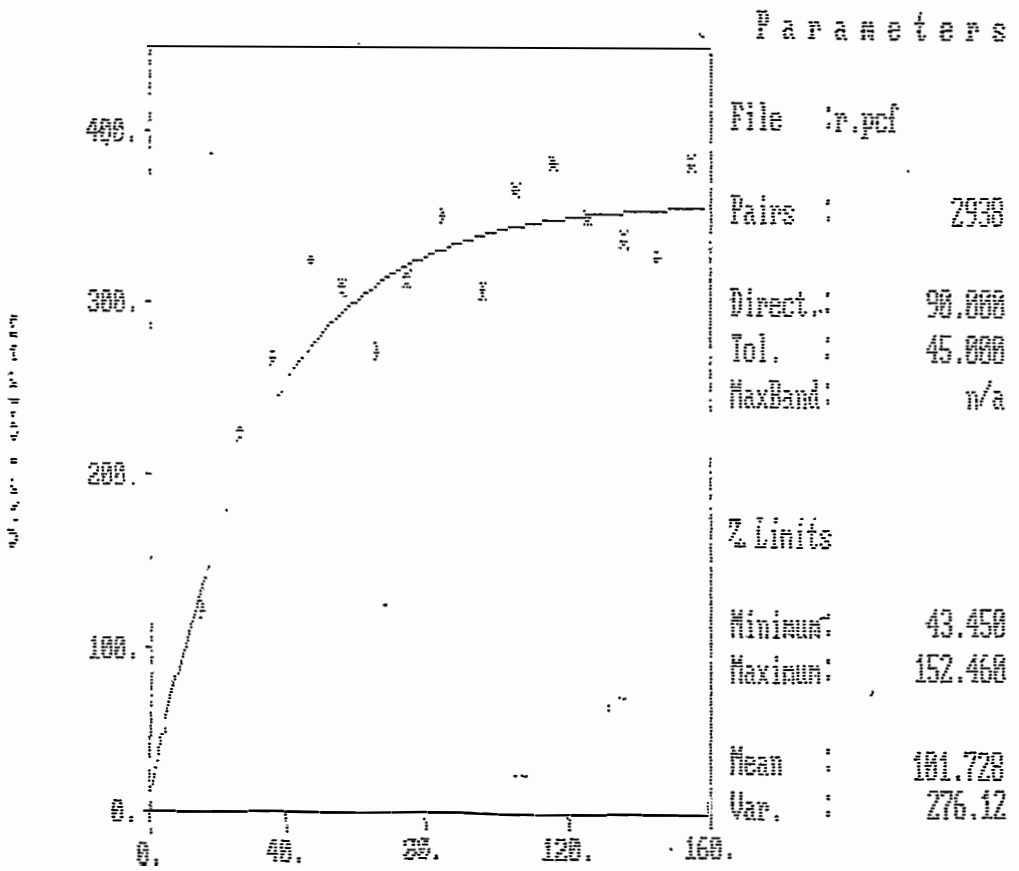


FIGURE 13

Description de la distribution des erreurs réduites sur les données de l'échantillon

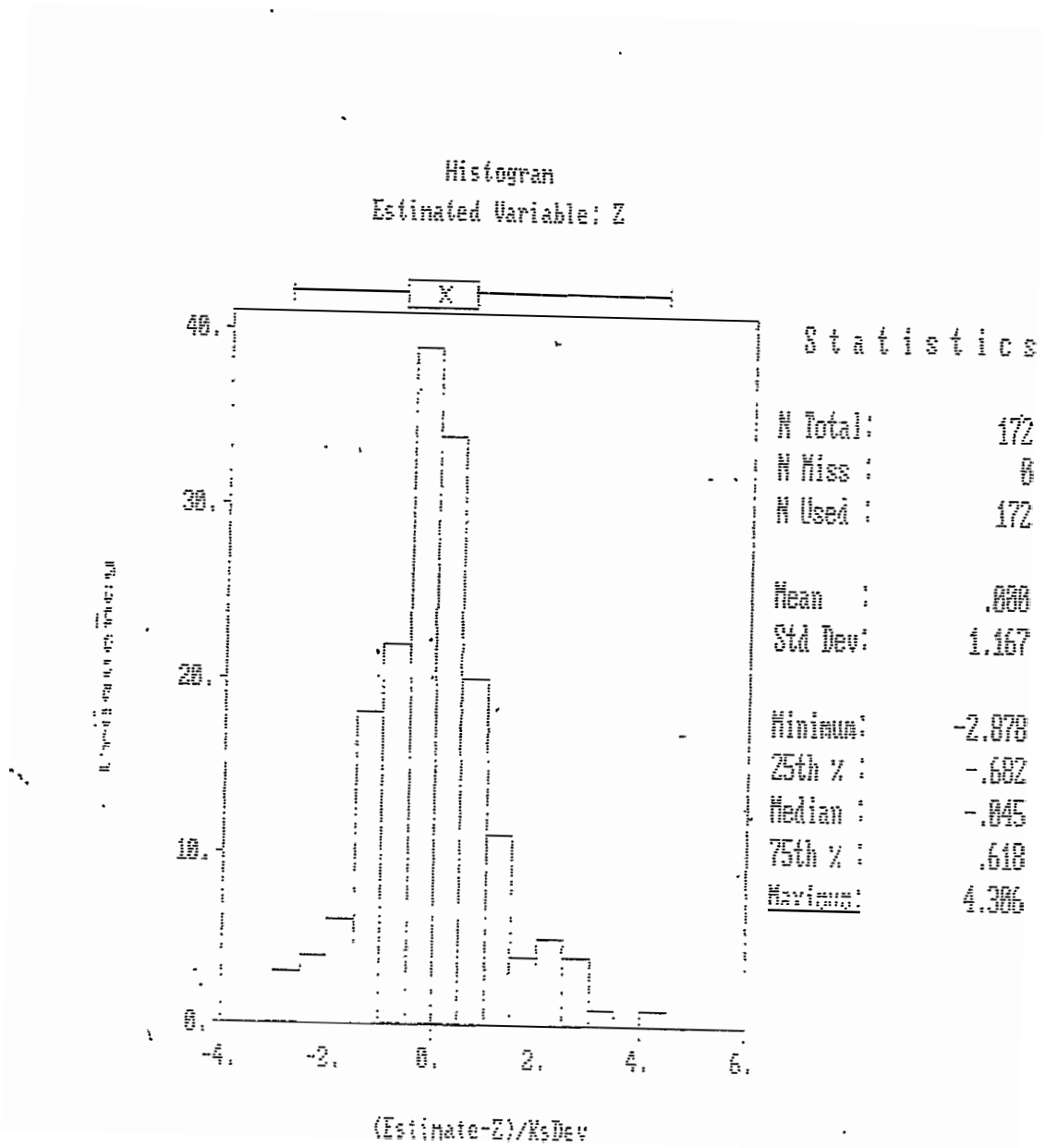


FIGURE 14 : Description de la distribution des erreurs réduites sur les individus hors échantillon

```
                                VALIDATION CROISEE : DEUXIEME PHASE              20:29 Sunday, September 5, 1993   4
                          COMPARAISON DE TOUTE LA GRILLE KRIGEE AVEC LA GRILLE VRAIE
                                          DONNEES DE jp1

                                        Univariate Procedure

Variable=SERRZ        Erreur de krigeage reduite

                Moments                                  Quantiles(Def=5)                                  Extremes

N                728  Sum Wgts        728      100% Max  7.915019      99%  6.466978        Lowest    Obs     Highest    Obs
Mean         0.14385  Sum        104.7229       75% Q3   0.758792      95%  2.578935      -3.57266(    417)  7.30261(    113)
Std Dev     1.513267  Variance   2.289976       50% Med  -0.04663      90%  1.728957      -3.50683(     29) 7.447123(    678)
Skewness    1.619094  Kurtosis   5.588649       25% Q1   -0.76576      10%  -1.44258      -3.16737(     10) 7.465905(    400)
USS         1679.877  CSS        1664.813        0% Min  -3.57266       5%  -1.83058      -2.94224(    115) 7.516295(    677)
CV          1051.974  Std Mean   0.056085                               1%  -2.81766      -2.88786(    378) 7.915019(    138)
T:Mean=0    2.564842  Prob>|T|     0.0105      Range     11.48767
Num ^= 0         728  Num > 0         354      Q3-Q1     1.524554
M(Sign)          -10  Prob>|M|     0.4813      Mode      -3.57266
Sgn Rank        1648  Prob>|S|     0.7718
W:Normal    0.889129  Prob<W       0.0001

                    Histogram                                  #  Boxplot
  7.75+*                                                       2     *
      .**                                                      4     *
      .*                                                       1     *
      .*                                                       3     *
      .*                                                       1     *
      .*                                                       2     0
      .*                                                       3     0
      .*                                                       2     0
      .**                                                      6     0
      .*                                                       2     0
      .*****                                                  13     |
      .*******                                                21     |
      .*********                                              27     |
      .********************                                   60     |
      .****************************                           83  +-----+
      .*********************************************         124  |  +  |
      .**************************************************    140  *-----*
      .***********************************                   105  +-----+
      .**********************                                 65     |
      .************                                           35     |
      .******                                                 17     |
      .***                                                     9     |
      .*                                                       1     0
  -3.75+*                                                      2     0
       ----+----+----+----+----+----+----+----+----+--
       * may represent up to 3 counts

                                                                 Normal Probability Plot
  7.75+                                                                                                  *
                                                                                                        **
                                                                                                         *
                                                                                                       **
                                                                                                        *
                                                                                                      *
                                                                                                     **
                                                                                                    *
                                                                                                  ***  +++
                                                                                                  *++++
                                                                                               ***
                                                                                            +***
                                                                                         +++***
                                                                                       ++*****
                                                                                    ++****
                                                                                 ******
                                                                            ******
                                                                        *****
                                                                     *****+
                                                                  ****+++
                                                               ****+++
                                                            *****+++
                                                            *++++
  -3.75+*
        +----+----+----+----+----+----+----+----+----+----+
            -2        -1         0        +1        +2
```

Annexes

A–1 La démonstration d'Efron

Nous voulons démontrer que $E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) \geq \gamma_{n-1}$.

L'idée de base est d'écrire $T(X_1, X_2, \ldots X_n)$ comme une somme de plusieurs fonctions de une, deux, trois, ... n variables, ces variables étant indépendantes.

Nous posons :

$$\begin{aligned}
\mu &= E\left(T\left(X_1, X_2, \ldots X_n\right)\right) = E(T) \\
A_i &= nE\left(T \mid X_i\right) \\
&\quad \text{fonction uniquement de } X_i \\
B_{ii'} &= n^2\left[E\left(T \mid X_i, X_{i'}\right) - E\left(T \mid X_i\right) - E\left(T \mid X_{i'}\right) + \mu\right] \\
&\quad \text{fonction uniquement de } X_i X_{i'} \\
N_{1,2,\ldots n} &= n^n\left[T - E\left(T \mid X_1 \ldots X_n\right) - \cdots - (-1)^n \mu\right]
\end{aligned}$$

Nous avons : $E(A_i) = E(B_{ii'}) = \cdots E(N_{1,2,\ldots n}) = 0$.
Les variables $A_i, B_{ii'}, \ldots$ sont non corrélées.

Enfin :

$$T_n = \mu + \frac{1}{n}\sum_i A_i + \frac{1}{n^2}\sum_{ii'} B_{ii'} + \cdots \frac{1}{n^n} N_{1,2,\ldots n}$$

Notons :

$$\sigma_A^2 = \mathrm{Var}(A_i) \quad \sigma_B^2 = \mathrm{Var}(B_{ii'}) \quad \text{etc}\ldots$$

et considérons la décomposition de T_n par rapport à T_{n-1}.

a) Nous pouvons considérer cette décomposition par rapport à $T_{(-i)}$ qui est construit à partir d'un $(n-1)$-échantillon.

$$T_{(-i)} = \mu^i + \frac{1}{n-1}\sum_{j \neq i} A_j^i + \frac{1}{(n-1)^2} \sum_{\substack{j<j' \\ j \neq i \text{ et } j' \neq i}} B_{jj'}^i + \cdots$$

Remarque :

$\mu^i = E\left(T_{(-i)}\right)$ est en fait l'espérance de l'estimateur T_{n-1} construit sur un $(n-1)$ -échantillon
$A_j^i = E\left(T\left(X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots X_n\right) \mid X_j\right) = E\left(T_{n-1} \mid X_j\right)$
etc ...
Ce qui signifie que μ^i, $A_j^i, \ldots$ sont indépendants de i et peuvent être notés μ, A_j (ces variables ne sont pas indépendantes de n).

Remarque :

$$\mathrm{Var}\left(T_{(-i)}\right) = \gamma_{n-1} = \frac{\sigma_A^2}{n-1} + C_{n-2}^1 \frac{\sigma_B^2}{2(n-1)^3} + C_{n-2}^2 \frac{\sigma_C^2}{3(n-1)^5} + \cdots$$

b) Si nous remarquons que :

$$\sum_i \left(T_{(-i)} - T_{(.)}\right)^2 = \frac{1}{n} \sum_{i<i'} \left(T_{(-i)} - T_{(-i')}\right)^2$$

Nous voyons :

$$E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) = \frac{1}{n} E\left(\sum_{i<i'} \left(T_{(-i)} - T_{(-i')}\right)^2\right) = \frac{n-1}{2} \mathrm{var}\left(T_{(-i)} - T_{(-i')}\right)$$

Nous allons chercher à exprimer $\left(T_{(-i)} - T_{(-i')}\right)$ en fonction des $A, B \ldots$ et à calculer $\mathrm{Var}\left(T_{(-i)} - T_{(-i')}\right)$ en fonction des $\sigma_A^2, \sigma_B^2, \ldots$ ce qui permettra de comparer $E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right)$ à γ_{n-1}.

c) Etudions le comportement des termes A_j dans la différence $\left(T_{(-i)} - T_{(-i')}\right)$.

Lorsque nous effectuons la différence $\left(T_{(-i)} - T_{(-i')}\right)$ il ne reste que les $A_{i'} - A_i$
De même pour les termes en B, il ne reste que $B_{ji} - B_{ji'}$ etc ...
D'où :

$$
\begin{aligned}
\left(T_{(-i)} - T_{(-i')}\right) &= \frac{1}{n-1}(A_{i'} - A_i) + \frac{1}{(n-1)^2}\sum(B_{ji'} - B_{ji}) + \cdots \\
\mathrm{Var}\left(T_{(-i)} - T_{(.)}\right) &= 2\left(\frac{1}{(n-1)^2}\sigma_A^2 + \frac{1}{(n-1)}C_{n-2}^1\sigma_B^2 + \frac{1}{(n-1)}C_{n-2}^2\sigma_C^2 + \cdots\right) \\
\text{et} & \\
E\left(\left(T_{(-i)} - T_{(.)}\right)^2\right) &= \frac{n-1}{2}\mathrm{Var}\left(T_{(-i)} - T_{(-i')}\right) \\
&= \frac{\sigma_A^2}{n-1} + \frac{C_{n-2}^1}{(n-1)^3}\sigma_B^2 + \frac{C_{n-2}^2}{(n-1)^5}\sigma_C^2 + \cdots
\end{aligned}
$$

d) Si nous comparons maintenant terme à terme :

$$E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) \quad \text{et} \quad \gamma_{\mathrm{n-1}}$$

Nous voyons que :

$$E\left(\sum_i \left(T_{(-i)} - T_{(.)}\right)^2\right) \geq \gamma_{\mathrm{n-1}}$$

et la différence est égale à :

$$\frac{1}{2}\frac{C_{n-2}^1}{(n-1)^3}\sigma_B^2 + \frac{2}{3}\frac{C_{n-2}^2}{(n-1)^4}\sigma_C^2 + \cdots$$

Remarque :

Les termes $\mu, \sigma_A^2, \sigma_B^2, \ldots$ sont des fonctions de n (hélas !).

Bibliographie

[1] B.M. Davis. Uses and abuses of cross-validation in geostatistics. *Mathematical Geology*, 19:241–248, 1987.

[2] Ministère de l'agriculture. Statistiques forestières en 1983. *Service Central des Enquêtes et Etudes Statistiques*, 30, Avril 1985.

[3] J. Durbin. A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46:477–480, 1959.

[4] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[5] E. Englund and A. Sparks. *GEO-EAS (Geostatistical Environmental assessment software) User's Guide.* U.S. Environmental Protection Agency, Las Vegas, 1988.

[6] H.L. Gray and W.R. Schucany. *The Generalized Jackknife Statistic.* Marcel Dekker, New-York, 1972.

[7] W. Hoeffding. A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.*, 19:293–325, 1948.

[8] J.J. Daudin, C. Duby and P. Trecourt. PCA analysis studied by the bootstrap and the infinitesimal jackknife method. *Statistics*, 20:255–270, 1988.

[9] J.J. Daudin, C. Duby and P. Trecourt. Stability of principal component analysis studied by the bootstrap method. *Statistics*, 19:241–258, 1988.

[10] M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Charles Griffin and Company limited, London, third edition, 1969.

[11] Rupert G. Miller. The Jackknife - a review. *Biometrika*, 61:1–15, 1974.

[12] F. Mosteller and J.W. Tukey. *Data Analysis and Regression.* Addison Wesley, 1977.

[13] M. Quenouille. Approximate tests of correlation in time series. *J. Roy. Statist. Soc. B*, 11:68–84, 1949.

[14] M. Quenouille. Notes on bias in estimation. *Biometrika*, 43:353–360, 1956.

[15] J.W. Tukey. Bias confidence in not quite large samples. *Ann. Math. Statist.*, 29, 1958. 614.

[16] K. M. Wolter. *Introduction to Variance Estimation.* Springer-Verlag, New-York, 1985.

Documents de travail de la mission biométrie

Déjà parus

1. LORIEUX M. Aspects statistiques de la cartographie des marqueurs moléculaires. 1994.

Centre
de coopération
internationale
en recherche
agronomique
pour le
développement

**Département
de gestion,
recherche,
documentation
et d'appui
technique**

Unité
de recherche
Biotrop

2477,
avenue du Val
de Montferrand
BP 5035
34032 Montpellier
Cedex 1
France