exploration interactive d'un nuage de points dans le plan

par Ph. Besse, A. de Falguerolles
*Universités de Toulouse II et III**

Dans le cadre de nos enseignements de statistique et d'informatique en première année du département statistique, études économiques et techniques quantitatives de gestion d'un Institut Universitaire de Technologie, il nous paraît intéressant de proposer aux étudiants de réaliser eux-mêmes un programme statistique à but pédagogique. Ce programme a des objectifs voisins de ceux d'un didacticiel décrit par A. Baccini, M.P. Martin et R. Moré (1984) : développer le "sens statistique" de l'utilisateur en lui permettant d'explorer les liens entre la forme d'un nuage de points dans le plan et les caractéristiques statistiques usuellement associées [1]. Cependant en demandant aux étudiants de produire eux-mêmes un tel programme, deux objectifs pédagogiques supplémentaires sont visés :
— familiarisation des étudiants avec les possibilités d'édition du matériel sur lequel ils travaillent (conception d'un éditeur interactif de nuage);
— sensibilisation des étudiants aux problèmes spécifiques du calcul statistique (calcul efficace de moyenne, de variance, de covariance...).

Cet article comprend deux parties. Dans la première partie on rappelle les problèmes bien connus que posent le calcul d'une variance, d'une covariance; on présente l'approche couramment retenue pour y répondre (calcul par récurrence). Il apparaît que les formules permettant ce calcul sont intéressantes tant d'un point de vue théorique que pratique. Dans la seconde partie, on présente le cahier des charges du logiciel demandé aux étudiants. On évoque certains problèmes techniques posés par le matériel sur lequel l'expérience a été tentée. Enfin on rend compte de l'utilisation de ce programme par nos étudiants.

(*) Laboratoire de Statistique et Probabilités, Université Paul Sabatier - U.A.-C.N.R.S. 745, 118, route de Narbonne, 31062 Toulouse cedex
et
Université de Toulouse - Le Mirail, I.U.T. (B), 5 allée A. Machado, 31058 Toulouse - Toulouse cedex.

I
Calcul statistique de moyenne, variance et covariance

On considère un caractère statistique quantitatif X défini sur une population Ω de taille finie $N = |\Omega|$. $X(i)$ désigne la valeur du caractère X observée sur l'unité statistique étiquetée i, $i \in \{1,\ldots,N\}$. Chaque unité statistique est munie d'un poids p_i, $p_i > 0$. Dans la pratique ces poids sont souvent égaux et de valeur 1. On note $\overline{X}$ et V^X respectivement la moyenne et la variance du caractère X :

$$\overline{X} = \frac{\sum_{i=1}^{N} p_i X(i)}{\sum_{i=1}^{N} p_i}$$

$$V^X = \frac{\sum_{i=1}^{N} p_i (X(i) - \overline{x})^2}{\sum_{i=1}^{N} p_i} = \frac{(\sum_{i=1}^{N} p_i)(\sum_{i=1}^{N} p_i X(i)^2) - (\sum_{i=1}^{N} p_i X(i))^2}{(\sum_{i=1}^{N} p_i)^2}$$

Si deux caractères quantitatifs X et Y sont observés conjointement sur cette même population, on note C^{XY} leur covariance :

$$C^{XY} = \frac{\sum_{i=1}^{N} p_i (X(i) - \overline{x})(Y(i) - \overline{y})}{\sum_{i=1}^{N} p_i}$$

Il vient encore :

$$C^{XY} = \frac{(\sum_{i=1}^{N} p_i)(\sum_{i=1}^{N} p_i X(i)Y(i)) - (\sum_{i=1}^{N} p_i X(i))(\sum_{i=1}^{N} p_i Y(i))}{(\sum_{i=1}^{N} p_i)^2}$$

1. Algorithmes élémentaires de calcul de moyenne et de variance

Il résulte des formules ci-dessus deux algorithmes de calcul que nous examinerons successivement. Leur extension au calcul de covariance est immédiat.

1.1. Algorithme 1

En fin d'algorithme, T donne la valeur de $\overline{x}$, S celle de V^x.

M : = 0
T : = 0
S : = 0

Pour $i = 1,\ldots,N$ *faire :*
    M : = M + p_i
    T : = T + $p_i \times X(i)$

Fin pour
T : = T/M

Pour $i = 1,\ldots,N$ *faire :*
    D : = $X(i)$ − T
    S : = S + $p_i \times D \times D$

Fin pour
S : = S/M

Remarque :

Cet algorithme est dit en deux passages car exige deux examens successifs de toutes les valeurs du caractère. Son utilisation manuelle est peu recommandée car toute erreur d'arrondi sur la valeur de la moyenne est portée dans le calcul de la variance.

1.2. Algorithme 2

En fin d'algorithme, T donne la valeur de $\overline{x}$, S celle de V^x.

M : = 0
T : = 0
S : = 0

Pour $i = 1,\ldots,N$ *faire :*
    M : = M + p_i
    T : = T + $p_i \times X(i)$
    S : = S + $p_i \times X(i) \times X(i)$

Fin pour
S : = (M × S − T × T)/(M × M)
T : = T/M

Remarque :

Cet algorithme est dit en un passage car n'exige qu'un seul examen des valeurs du caractère. Appelé "textbook algorithm" par les anglo-saxons, il est couramment utilisé dans les calculs à la main car donne des résultats exacts. Il est aussi implanté sur la plupart des calculettes à fonctions statistiques car n'exige que trois registres. Il présente clairement des risques de dépassement de capacité puisqu'on calcule une somme de carrés. Dans une conduite manuelle il conduit le cas échéant à manipuler de

"grands chiffres". Aussi fait-on souvent une transformation préalable des données.

1.3. Problèmes posés par l'emploi de ces algorithmes sur ordinateur.

L'inconvénient majeur de l'algorithme 1 est qu'il exige soit deux lectures de la même série statistique, soit une lecture et un stockage en mémoire de cette série. Il est cependant précis [2].

L'emploi de l'algorithme 2 doit être prohibé pour de grandes séries statistiques malgré son intérêt de n'exiger qu'un seul examen de la série statistique. En effet, même si aucun dépassement de capacité ne se produit, l'expression

$$S := (M \times S - T \times T)/(M \times M)$$

peut déterminer des résultats assez fantaisistes. Ce risque est lié à la représentation sous forme flottante des nombres et dépend du nombre de chiffres caractéristiques retenus.

Illustrons cet effet sur un exemple caricatural. Soit à calculer la moyenne et la variante de deux nombres 9999 et 9998 sur une machine ne retenant que quatre chiffres significatifs et tronquant les résultats des opérations. En fin d'exécution du bloc *pour*, on a :

$$M = 0{,}2000.10^{1} \text{ (d'où } M \times M = 0{,}4000.10^{1})$$
$$T = 0{,}1999.10^{5} \text{ (d'où } T \times T = 0{,}3996.10^{9})$$
$$S = 0{,}1999.10^{9} \text{ (d'où } M \times S = 0{,}3998.10^{9})$$

L'expression $S = (M \times S - T \times T)/(M \times M)$ donne :

$$\frac{(0{,}3998.10^{9} - 0{,}3996.10^{9})}{0{,}4000.10^{1}} \text{ soit } \frac{0{,}2000.10^{6}}{0{,}4000.10^{1}} \text{ donc } 0{,}5000.10^{5}.$$

L'expression T : T/M donne $0{,}9995.10^{4}$.

On voit que sur une telle machine on obtient une moyenne égale à 9995 (valeur exacte 9998,5 !) et une variance égale à $0{,}5.10^{5}$ (valeur exacte 0,25 !!). Il est à noter que l'algorithme 1 aurait donné une variance égale à 12,5.

2. Calcul par récurrence

Ce calcul conduit à un algorithme qui combine les avantages des deux algorithmes précédents : un seul passage et bonnes propriétés d'erreur.

2.1. Formules de récurrence.

L'idée de la méthode est simple. On note respectivement $\overline{x}_n$ et V_n^X la moyenne et la variance de la restriction du caractère X à n unités statistiques ($1 \leqslant n < N$); M_n désigne la somme des poids des n unités statistiques concernées. On considère une $n+1$ ième unité statistique de

poids p_{n+1} et présentant la valeur $X(n+1)$. On exprime alors $\overline{x}_{n+1}$ et V^X_{n+1} en fonction de $\overline{x}_n$, V^X_n, M_n, p_{n+1} et $X(n+1)$.

Il vient : $$\overline{x}_{n+1} = \frac{M_n}{M_{n+1}}\overline{x}_n + \frac{p_{n+1}}{M_{n+1}}X(n+1)\,.$$

Ainsi : $$\overline{x}_{n+1} - \overline{x}_n = \frac{p_{n+1}}{M_{n+1}}(X(n+1) - \overline{x}_n) = \frac{p_{n+1}}{M_n}(X(n+1) - \overline{x}_{n+1})\,.$$

De plus : $$V^X_{n+1} = \frac{M_n}{M_{n+1}}V^X_n + \frac{p_{n+1}M_n}{M^2_{n+1}}(X(n+1) - \overline{x}_n)^2$$

Soit encore :

$$V^X_{n+1} = \frac{M_n}{M_{n+1}}V^X_n + \frac{p_{n+1}}{M_n}(X(n+1)-\overline{x}_{n+1})^2$$

$$= \frac{M_n}{M_{n+1}}V^X_n + \frac{M_n}{p_{n+1}}(\overline{x}_{n+1}-\overline{x}_n)^2$$

Ces formules suggèrent un algorithme réalisant un compromis entre les algorithmes 1 et 2 : un seul passage ; une diminution des risques d'arrondis puisqu'élevant au carré soit des écarts à la moyenne courante ($X(n+1)-\overline{x}_n$ ou $X(n+1)-\overline{x}_{n+1}$) soit des écarts entre moyennes courantes ($\overline{x}_{n+1}-\overline{x}_n$).

Outre leur intérêt algorithmique ces formules montrent comment chaque observation affecte la valeur de la moyenne ou de la variance.

2.2. Analogie avec le lissage exponentiel

Lorsque les valeurs $X(1),\ldots,X(N)$ sont celles d'une série chronologique, il est assez courant de procéder à une estimation glissante de la moyenne par la formule suivante : $\overline{x}_{n+1} = \overline{x}_n + \alpha(X(n+1)-\overline{x}_n)$ où α est un nombre compris entre 0 et 1 choisi a priori (par exemple $\alpha = 0{,}05$).

L'analogie avec la formule de récurrence pour la moyenne est claire. Le lissage exponentiel revient à donner des poids croissants aux observations :

$$p_1, p_2 = \frac{\alpha}{1-\alpha}p_1, \ldots, p_n = \frac{\alpha}{(1-\alpha)^{n-1}}p_1, \ldots$$

L'estimation glissante de la variance est alors donnée par l'expression :

$$V_{n+1} = (1-\alpha)\,[V_n + \alpha(X(n+1)-\overline{x}_n)^2]$$

2.3. Application à la statistique mathématique.

Récemment S.M. Stigler (1984) a rappelé comment ces formules peuvent être utilisées en statistique mathématique pour établir les propriétés des estimateurs usuels de la moyenne et de la variance d'un échantillon de variables aléatoires indépendantes et de mêmes lois normales $N(\mu,\sigma^2)$.

2.4. Algorithme 3

Différentes formulations peuvent être retenues compte tenu des différentes formes des expressions de récurrence. Suivant l'étude de E.A. Young et E.M. Cramer (1971) on retient l'algorithme suivant :

$M := p_1$
$T := p_1 \times X(1)$
$S := 0$
Pour $i = 2,\ldots,N$ *faire*
$M_0 := M$
$M := M + p_i$
$T := T + p_i \times X(i)$
$D := M \times X(i) - T$
$S := S + (p_i \times D \times D)/(M \times M_0)$
Fin pour
$T := T/M$
$S := S/M$

Ces auteurs ont montré que cet algorithme a des propriétés d'erreur analogues à celles de l'algorithme 1.

3. Application au calcul du coefficient de corrélation linéaire de Bravais-Pearson et à la régression linéaire simple.

Considérons le cas où l'on observe conjointement deux caractères quantitatifs X et Y. Les algorithmes 1,2, et 3 s'adaptent aisément au calcul de la covariance. En particulier, on a les formules suivantes de récurrence

$$C_{n+1}^{XY} = \frac{M_n}{M_{n+1}} C_n^{XY} + \frac{M_n p_{n+1}}{M_{n+1}^2} (X(n+1) - \overline{x}_n)\,(Y(n+1) - \overline{y}_n)$$

ou encore :

$$C_{n+1}^{XY} = \frac{M_n}{M_{n+1}} C_n^{XY} + \frac{p_{n+1}}{M_n} (X(n+1) - \overline{x}_{n+1})\,(Y(n+1) - \overline{y}_{n+1})$$

Il en résulte une formule de récurrence pour le coefficient de corrélation linéaire de Bravais-Pearson $r^{XY} = C^{XY} / \sqrt{V^X V^Y}$:

$$r_{n+1}^{XY} = \frac{r_n^{XY} + \frac{p_{n+1}}{M_{n+1}} \left(\frac{X(n+1) - \overline{x}_n}{\sqrt{V_n^X}}\right)\left(\frac{Y(n+1) - \overline{y}_n}{\sqrt{V_n^Y}}\right)}{\left\{1 + \frac{p_{n+1}}{M_{n+1}} \left[\frac{X(n+1) - \overline{x}_n}{\sqrt{V_n^X}}\right]^2\right\}^{\frac{1}{2}} \left\{1 + \frac{p_{n+1}}{M_{n+1}} \left[\frac{Y(n+1) - \overline{y}_n}{\sqrt{V_n^Y}}\right]^2\right\}^{\frac{1}{2}}}$$

On peut encore considérer la régression de Y sur X : $Y = aX + b$. Il vient alors :

$$a_{n+1} = \frac{a_n + \frac{p_{n+1}}{M_{n+1}} \left(\frac{X(n+1) - \overline{x}_n}{\sqrt{V_n^X}} \right) \left(\frac{Y(n+1) - \overline{y}_n}{\sqrt{V_n^Y}} \right) \sqrt{\frac{V_n^Y}{V_n^X}}}{1 + \frac{p_{n+1}}{M_{n+1}} \left(\frac{X(n+1) - x_n}{\sqrt{V_n^X}} \right)^2}$$

Les deux formules ci-dessus montrent encore comment chaque observation affecte la valeur du coefficient de corrélation linéaire et de la pente de la droite de régression :

1) par son poids relatif ($\frac{p_{n+1}}{M_{n+1}}$) ,

2) par sa position relative dans le nuage ($\frac{X(n+1) - \overline{x}_n}{\sqrt{V_n^X}}$ et $\frac{Y(n+1) - \overline{y}_n}{\sqrt{V_n^Y}}$).

4. Suppression d'une unité statistique

Les formules de récurrence permettent aussi d'étudier les effets de la suppression d'une unité statistique. Il suffit de considérer cette unité comme une unité statistique entrante de poids égal à l'opposé de son poids initial.

II
Logiciel d'Etude Interactive de Nuage

Ce logiciel a pour objet de permettre de dessiner un nuage de points et de suivre les effets de l'adjonction ou de la suppression de points sur les valeurs des caractéristiques statistiques usuellement associées à ce nuage.

1. Cahier des charges

Les étudiants doivent réaliser un éditeur de nuage de points répondant aux caractéristiques suivantes.

L'écran du terminal ou du micro-ordinateur est divisé en deux zones. La première contient un nuage de points plan dans un repère orthonormé ; la deuxième les caractéristiques du nuage (nom du nuage, nombre de points, moyenne des x , moyenne des y , variances, écarts-types, covariance, coefficient de corrélation linéaire, paramètres de la droite de régression).

Le principe d'utilisation est le même que celui d'un éditeur de texte en mode page ; l'utilisateur déplace un curseur sur l'écran et, à volonté, ajoute ou élimine un point à l'emplacement de celui-ci. Les caractéristiques du nuage sont remises à jour à chaque modification.

D'autres commandes doivent être disponibles :

— édition du nuage après centrage des variables,

— édition du nuage après centrage et réduction des variables,

— retour au nuage initial,

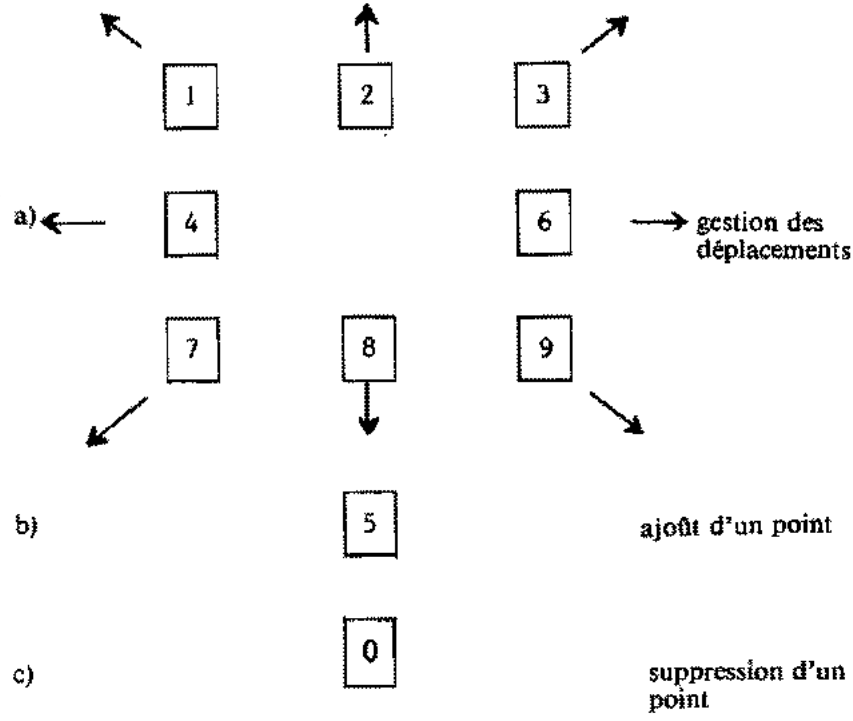
— sauvegarde sur disque des coordonnées des points et des caractéristiques du nuage,

— retour au menu principal proposant quatre choix (création, modification, suppression d'un nuage, arrêt du programme).

2. Remarques techniques

L'intérêt des formules de récurrence est évident. De plus elles donnent une vitesse d'exécution et un temps de réponse satisfaisant même sur une petite machine.

Ce logiciel étant réalisé en BASIC sur un SOLAR 16/65, matériel qui n'a pas la souplesse d'un micro-ordinateur, le problème principal réside dans la gestion du déplacement du curseur. En particulier il n'est pas possible d'interroger du programme la console pour connaître la position du curseur ; c'est à l'utilisateur de la gérer par l'intermédiaire, par exemple, du clavier numérique :



Cette gestion rend nécessaire un "retour charriot" à chaque déplacement ou modification. Temps d'exécution et confort d'utilisation en sont affectés.

3. Intérêt pédagogique

L'intérêt doit être évalué à deux niveaux : réalisation du logiciel et utilisation pédagogique.

Le sujet, suffisamment attractif, a suscité une bonne motivation chez les étudiants qui ont développé des initiatives intéressantes sur le plan informatique. De plus certains étudiants ont saisi l'intérêt et la spécificité du calcul statistique sur ordinateur.

Il nous paraissait aussi qu'à travers son utilisation, ce logiciel pouvait contribuer à la formation d'un certain sens statistique chez ses utilisateurs.

En explorant les liens entre forme d'un nuage et caractéristiques numériques associées, l'utilisateur de cet éditeur devait pouvoir notamment se familiariser avec les signes et les ordres de grandeur des coefficients de corrélation linéaire et découvrir des cas de fausse corrélation. En fait deux produits ont été présentés en libre accès à la sagacité des étudiants : le didacticiel "nuage" développé par A. Baccini et al. (1984) et leur programme. "Nuage", très directif, invitait l'utilisateur à deviner les valeurs des caractéristiques du nuage proposé tandis que l'éditeur précédent laissait plus de liberté pour l'exploration de celles-ci. A l'expérience on a constaté qu'après mise au point des programmes la première attitude des étudiants a surtout consisté à faire des "dessins" sur l'écran.

Il semble donc que, pour conduire à l'apprentissage souhaité, l'éditeur nécessite la présence permanente d'un enseignant pour lequel il constitue un bon outil pédagogique. Par contre "nuage", conçu a priori comme didacticiel, est d'un emploi parfaitement autonome : le rôle de l'enseignant y est intégré par l'aspect directif de son déroulement.

Références :

[1] BACCINI A., MARTIN M.P., et MORÉ R. : *"E.A.O. en statistique"*, bulletin de l'A.P.M.E.P., 343 (1984),215-224.

[2] CHAN T.F., GOLUB G.H., and Le VEQUE R.J. : *"Updating formulae and a pairwise algorithm for computing sample variances"*, in COMPSTAT 82, ed. H. Caussinus, P. Ettinger and R. Tomassone, Physica-Verlag, Wien 1982.

[3] SEARLE S.R. : *"The recurrence formulae for means and variances"*, Teaching Statistics, Vol. 5, N° 1, (1983), 7-10.

[4] STIGLER S.M. : *"Kruskal's proof of the joint distribution of $\overline{x}$ and s^2"*, The American Statistician, Vol. 38, N° 2, (1984) 134-135.

[5] YOUNGS E.A., and CRAMER E.M. : *"some results relevant to choice of sum and sum-of-product algorithms"*, Technometrics 13 (1971) 657-665.