échanges

vive le changement !

par Bernard Parzysz

Au vu de certains exercices de Statistiques posés récemment au Baccalauréat(*), il me semble bien qu'une notion est nettement sous-employée : je veux parler du changement de variable. Prenons par exemple le sujet des académies "parisiennes" en septembre 1984 : il s'agissait en l'occurence d'étudier la corrélation entre la note obtenue à l'écrit (x_i) et celle obtenue à l'oral (y_i) par huit élèves, lors des épreuves anticipées de français. La troisième question de cet exercice demandait de remplir le tableau suivant :

x_i	y_i	x_i^2	x_iy_i	y_i^2
8	13	?	?	?
4	9	?	?	?

(*N.B.* : Sont seules indiquées ici la première et la dernière ligne du tableau. Nous n'épiloguerons pas sur le cas des malheureux élèves qui — ayant mal interprété les pointillés — ont cru qu'il suffisait de remplir ces deux lignes).

(*) En particulier : Paris B juin 1984 - Lyon D juin 1984 - Paris B septembre 1984 - Amiens B juin 1985.

La question impose donc de *ne pas* faire de changement de variable. Or, lequel d'entre nous, lorsqu'il veut calculer ("à la main") la moyenne des notes qu'a obtenues une classe à un devoir, par exemple, le fait "brutalement" en additionnant les dites notes, alors qu'il est incomparablement plus simple de ne considérer que les différences par rapport à 10 (ce qui revient donc à translater les notes de 10 vers le bas) ?

On pourra arguer :

1° - qu'il y a, dans le cas qui nous occupe ici, peu de nombres ;
2° - que les calculatrices vont aussi vite — du moins à notre échelle — pour calculer le carré de 3784 que celui de 3.

A quoi je répondrai :

1° - que l'on a parfois des séries importantes à traiter. Plus précisément : avec des séries à gros effectif de nombres ayant beaucoup de chiffres significatifs, on arrive parfois à dépasser les capacités d'affichage des calculatrices courantes (pour Σx_i^2), ce qui conduit à remplacer une valeur exacte par une valeur approchée involontaire (notation scientifique) ;
2° - que l'on n'a pas toujours une calculatrice à sa disposition ; et surtout
3° - qu'il s'agit d'une question de principe.

La technique du changement de variable me paraît en effet fondamentale à faire acquérir, que ce soit en Statistiques ou ailleurs (ceux qui, par exemple, ont eu à faire tracer des courbes par un ordinateur ne me contrediront pas).

On rétorquera peut-être que démontrer des formules telles que $\overline{X+k}=\overline{X}+k$, $\overline{k.X}=k.\overline{X}$, $\sigma(X+k)=\sigma(X)$ et $\sigma(k.X)=|k|.\sigma(X)$ est bien difficile pour des élèves non scientifiques (utilisation des propriétés de Σx_i^2). Mais de véritables démonstrations peuvent être faites sans utiliser aucun calcul : il suffit que les élèves aient bien compris ce que sont la moyenne et la dispersion d'une série statistique. A partir de là, et en prenant au besoin un exemple tel que celui des notes d'une classe, déjà évoqué plus haut, on peut établir les résultats. Ainsi :

a) si j'augmente chaque note de 2 points (acclamations), la moyenne de la classe va être augmentée de 2 points, la dispersion ne changeant pas puisque les écarts à la moyenne restent les mêmes ;

b) si je divise toutes les notes par 2 (huées), c'est-à-dire si je les multiplie par 0,5, la moyenne de la classe est, elle aussi, divisée par 2, ainsi que l'écart-type (les écarts étant cette fois divisés par 2) ;

c) pour ce qui est de la covariance on pourra, de façon analogue, voir comment varie le produit $(x_i-X)(y_i-Y)$.

Enfin, à titre d'exercice, on pourra faire démontrer, à partir des résultats précédents, que le coefficient de corrélation reste invariant dans les

changements affines de variable du type $X' = aX + b$, $Y' = cY + d$, avec a et c positifs. On pourra également donner de petites "astuces", du genre :

• choisir comme nouvelle "origine" de la variable une valeur proche (au jugé) de la moyenne, valeur *effectivement prise* par la variable (ce qui fournit des zéros dans la ligne correspondante du tableau) ;

• dans le cas d'une série double, ne pas choisir les nouvelles "origines" dans la même ligne (d'où *deux* lignes contenant des zéros)...

Exemple : Soit la série double ci-dessous (indiquant une production en fonction de l'année), dont on veut calculer le coefficient de corrélation linéaire :

X	1974	1976	1978	1980	1982	1984
Y	114,8	117,0	119,8	121,2	125,6	128,2

Premier cas : calcul direct :

X	Y	X^2	XY	Y^2
1974	114,8	3896676	226615,2	13179,04
1976	117,0	3904576	231192,0	13689,00
1978	119,8	3912484	236964,4	14352,04
1980	121,2	3920400	239976,0	14689,44
1982	125,6	3928324	248939,2	15775,36
1984	128,2	3936256	254348,8	16435,24
Somme 11874	726,6	23498716	1438035,6	88120,12

On a alors :

$$\overline{X} = \frac{11874}{6} = 1979$$

$$\overline{Y} = \frac{726,6}{6} = 121,1$$

$$\sigma^2(X) = \frac{23498716}{6} - \overline{X}^2 = \frac{35}{3}$$

$$\sigma^2(Y) = \frac{88120,12}{6} - \overline{Y}^2 = \frac{6443}{300}$$

$$\mathrm{Cov}(X, Y) = \frac{1438035,6}{6} - \overline{X}.\overline{Y} = 15,7$$

(Ce qui donne, pour le coefficient de corrélation, environ 0,992. Mais là n'est pas la question...).

Second cas : On prend cette fois comme nouvelle "origine" pour l'année 1980, et on divise par 2 (soit : $x' = \frac{X - 1980}{2}$).

De même, on prend comme nouvelle origine pour Y la valeur 119,8, et on multiplie par 5 (soit : $Y' = 5.(Y - 119{,}8)$).

D'où le tableau suivant :

	X'	Y'	X'^2	$X'Y'$	Y'^2
	−3	−25	9	75	625
	−2	−14	4	28	196
	−1	0	1	0	0
	0	7	0	0	49
	1	29	1	29	841
	2	42	4	84	1764
Somme	−3	39	19	216	3475

On a maintenant :

$$\overline{X'} = -\frac{3}{6} = -\frac{1}{2}$$

$$\overline{Y'} = \frac{39}{6} = \frac{13}{2}$$

$$\sigma^2(X') = \frac{19}{6} - \overline{X'}^2 = \frac{35}{12}$$

$$\sigma^2(Y') = \frac{3475}{6} - \overline{Y'}^2 = \frac{6443}{12}$$

$$\mathrm{Cov}(X', Y') = \frac{216}{6} - \overline{X'}.\overline{Y'} = \frac{157}{4}$$

On peut alors calculer le coefficient de corrélation, et on trouve — bien sûr — la même valeur que plus haut.

La comparaison des deux tableaux est suffisamment éloquente pour que je n'y insiste pas. Certes, la calculatrice est un outil irremplaçable ; mais il ne faudrait pas qu'elle serve d'alibi à une certaine forme de paresse intellectuelle.

Il est vrai que le changement de variable n'est pas une pratique inconnue des sujets du Baccalauréat. Ainsi, lorsque l'une des deux variables "naturelles" est l'année, on lui substitue souvent le "rang" de l'année.

Ce n'est peut-être pas le meilleur choix possible (voir plus haut : il vaut mieux avoir une variable qui prend les valeurs entières de -3 à 2, plutôt qu'une variable qui prend les valeurs entières de 1 à 6). Alors, pourquoi l'imposer ? Croit-on que les candidats sont incapables de trouver seuls un bon changement de variable ? Ou bien, plus prosaïquement, cette façon de faire a-t-elle pour but de faciliter la tâche des correcteurs, tout le monde utilisant les mêmes variables ?

J'irai même plus loin : quels objectifs veut-on tester à l'examen dans un exercice de Statistiques ? Pratiquement tous portent sur l'ajustement linéaire par la méthode des moindres carrés. S'il s'agit d'évaluer les connaissances des candidats, alors pourquoi leur donner les formules, telles celles de l'équation cartésienne de la droite de régression(**) ? D'autant plus que ces formules risquent fort de ne pas correspondre, dans la forme, à celles qu'ils ont vues en classe, et par conséquent de les perturber. Ainsi, le coefficient directeur de la droite de régression de Y en X peut être donné sous la forme

$$\frac{\Sigma(x_i-X)(y_i-Y)}{\Sigma(x_i-X)^2},$$

ou bien sous la forme $\dfrac{\mathrm{Cov}(X,Y)}{\sigma^2(X)}$,

avec $\mathrm{Cov}(X,Y)=\dfrac{1}{n}\ \Sigma(x_i-X)(y_i-Y)$ et $\sigma^2(X)=\dfrac{1}{n}\ \Sigma(x_i-X)^2$

(entre autres).

S'il s'agit de tester les aptitudes à conduire une étude statistique, alors pourquoi imposer un changement de variable (ou un non-changement de variable) ? Dans de tels exercices, on ne laisse plus aucune initiative à l'élève : il lui suffit de savoir que la droite de régression peut servir, dans certains cas, à estimer une évolution, et de savoir se servir de sa calculatrice (les formules et même la présentation du tableau étant données). Ce n'est finalement que ce dernier point que l'on peut évaluer dans de telles conditions.

Serait-il irréaliste de proposer aux candidats de faire preuve d'esprit critique (validité d'un ajustement), et de montrer leur capacité à utiliser les données de façon à ne faire qu'un minimum de calculs et à en tirer le meilleur parti possible (le meilleur rapport qualité/prix, en quelque sorte) ? Et, à la limite, puisqu'il ne s'agit en fait que d'un "exercice de style" sur des données très réduites, et non de l'étude d'une "vraie" série statistique, pourquoi ne pas inciter les candidats à utiliser le moins possible leur calculatrice pour cette épreuve ? Voilà qui pourrait peut-être donner un regain d'intérêt à un type d'exercice par ailleurs passablement stéréotypé.

(**) Paris B juin 1985.