drôle d'exemple

par Bernard Parzysz

Le coefficient de corrélation linéaire de deux variables statistiques se conduit parfois de façon tout à fait étrange, pour ne pas dire inconvenante.

Ainsi, prenons par exemple la série statistique double suivante, indiquant une production annuelle :

X	1981	1982	1983	1984	1985
Y	72	101	141	185	275

La vue du nuage de points n'incite guère à faire un ajustement linéaire ; on pense plutôt à une croissance de type exponentiel. D'où l'idée de prendre le logarithme népérien (que nous noterons Z) de la variable Y.

Mettons-nous maintenant à la place de l'élève qui se trouve confronté (en Terminale B, par exemple) au problème de trouver le coefficient de corrélation linéaire des variables X et Z. Sa calculatrice fonctionne parfaitement ; elle est de type "scientifique" (quoique sans fonctions statistiques), donc tout va bien.

"Mais — se dit notre élève — combien de décimales vais-je prendre ?" ... ET C'EST ICI QUE SES ENNUIS RISQUENT DE COMMENCER.

Supposons en effet qu'il décide de s'en tenir à n chiffres après la virgule dans tous les calculs. Supposons également qu'il arrondisse — c'est un raffiné — "au plus près", c'est-à-dire par défaut si le $(n+1)^e$ chiffre est inférieur ou égal à 4, et par excès dans le cas contraire. (Notons d'ailleurs que le problème n'est pas, pour lui, de savoir si ces chiffres sont significatifs (comment le pourrait-il, d'ailleurs ?) ; il se donne simplement une règle de gestion de ses décimales).

Voyons donc ce qu'il va trouver comme résultat pour le coefficient de corrélation, suivant la valeur qu'il aura prise pour n. Le tableau ci-dessous donne les valeurs r_n trouvées pour le coefficient, en fonction de n :

n	2	3	4	5	6
r_n	1,04	1,007	0,9995	0,99883	0,998738

(*N.B. : Nous n'avons pas fait figurer le cas $n=1$, que d'aucuns pourraient* trouver trop caricatural).

Comme on le voit, notre héros malheureux (et d'autant plus malheureux qu'il connaît bien son cours) va se trouver bien embarrassé, s'il choisit de ne prendre que deux ou trois chiffres après la virgule, avec un coefficient de corrélation supérieur à 1 !(*)

Que va-t-il penser ?

1° - "Il y a une erreur d'énoncé". Puis, aussitôt : "Mais non, puisque pour toute distribution, le coefficient de corrélation est au plus égal à 1". Et il en viendra au

2° - "Je me suis trompé dans les calculs". Et il recommencera tous ses calculs, ce qui sera bien entendu sans espoir s'il ne pense pas au

3° - "Je n'ai pas pris suffisamment de chiffres après la virgule".

... Mais le pensera-il vraiment, sauf s'il s'est déjà fait "avoir" de la même façon ?

Nous n'osons pas imaginer que cette mésaventure puisse lui arriver au cours d'un contrôle en classe, ou même — mais c'est heureusement tout à fait exclu — au cours d'un examen.

ALORS ?

Vite, vite, hâtons-nous de "piéger" nous-mêmes nos élèves pour éviter qu'ils ne se piègent eux-mêmes par la suite, à un moment où l'enjeu serait d'importance. C'est peut-être un service à leur rendre.

(*) Bien sûr, pour nous qui, au temps où les calculatrices n'existaient pas, avions la "chance" d'utiliser des tables de logarithmes à 5 décimales, il n'était pas question, à l'époque, de nous contenter de si peu. Mais il n'en est certainement pas de même pour l'élève à qui l'on a répété maintes fois que $\ln 2$ vaut environ 0,69, que $\ln 3$ vaut environ 1,1, et qui a toujours eu bien assez d'une ou deux décimales pour placer des points sur un graphique, lors d'une construction de courbe.