A propos de l'enseignement de la statistique en collège et au lycée

C. Robert

Dans cet atelier, qui, étant donné le nombre de participants (environ 80) était plutôt une conférence, j'ai présenté quelques problèmes généraux de l'enseignement de la statistique, qui ont été illustrés principalement par le cas de l'ajustement linéaire dont on fait un grand usage dans l'enseignement secondaire.

D'où la présentation en deux paragraphes de ce «compte-rendu d'atelier».

I - Quelques erreurs fréquentes de l'enseignement de la statistique

La statistique enseignée dans le secondaire voudrait être de la statistique descriptive mais consiste le plus souvent en une fastidieuse série d'exercices de calculs de moyenne, d'écart-types et de tracés d'histogrammes - c'est à peu près aussi intéressant que de lire un annuaire du téléphone sans aucune raison de le faire.

Il y a une drôle de méthode pour soi-disant calculer cette moyenne, méthode repérée par le terme de «la convention usuelle ». Si on regroupe les données en classes, si on affecte les données d'une classe au centre de celle-ci et si on calcule la moyenne des données ainsi construites, le nombre obtenu est une bonne approximation de la moyenne des données originelles mais ce n'est pas la moyenne elle-même. Cette méthode de calcul n'a aucune raison

d'être pratiquée si on dispose des données et pas simplement des effectifs des classes de regroupement.

D'autre part, on aborde en première les probabilités, donc les tirages de cartes, de boules coloriées dans des urnes imaginaires, de lancers de dés et de pièces parfaitement honnêtes etc. On arrive ainsi à une étape où une réflexion sur la modélisation devient inévitable. On ne peut pas dire, comme le font certains ouvrages, qu'un modèle d'une expérience aléatoire est une expérience de tirages de boules dans une urne qui conduit au même espace probabilisé que l'expérience initiale (pourquoi, dans le champ de la statistique, un dispositif expérimental serait-il appelé modèle alors qu'en physique un modèle est toujours un objet théorique ?). La notion de probabilité d'un événement est le plus souvent définie comme limite de distribution empirique, et le fait que cette limite théorique n'a de sens que dans un modèle est escamoté aux yeux des élèves. (Cette situation perdure ultérieurement chez ceux qui entendent parler de la loi des grands nombres sans en faire la démonstration : ils pensent alors naturellement qu'il s'agit uniquement d'une loi empirique -puisqu'en mathématiques, ce qui se démontre s'appelle un théorème-).

Que les enseignants du secondaire se consolent (?): l'enseignement de la statistique en DEUG n'est pas dépourvu d' incohérences. Ainsi, dans le cadre de tests d'hypothèses notamment, on parle de la "vraie valeur" de la probabilité qu'une pièce tombe sur le coté pile, ce qui semblerait dire qu'il y a un vrai modèle et c'est là aussi en contradiction avec les autres sciences.

Le cas de l'enseignement de la statistique n'est cependant pas aussi grave qu'on pourrait le croire, dans la mesure ou les incohérences qu'il véhicule sont, comme on vient de le voir, bien visibles et de fait on peut les corriger facilement. Et si les enseignants ont plus facilement accès à des logiciels simples de statistique ou à des tableurs style Excel, alors l'enseignement des statistiques sera même plaisant....

2 - A propos de l'ajustement linéaire

2.1.Si on veut exprimer le fait qu'un nuage de points N est «allongé» grâce à une droite, comment choisir cette droite ? Si $N=\{ M_i, i=1..,n \}$, notons ε_i, ε'_i, ε''_i les longueurs respectives des segments $[M_iN_i]$, $[M_iM'_i]$,$[M_iM''_i]$ (cf. figure 1) et x_i et y_i les coordonnées de M_i. On peut choisir par exemple :

1 - une droite qui minimise $\varepsilon_1+...+\varepsilon_n$.

2 - une droite qui minimise $\varepsilon'_1+...+\varepsilon'_n$.

3 - une droite qui minimise $\varepsilon''_1+...+\varepsilon''_n$.

4 - une droite qui minimise $\varepsilon_1^2 + \ldots + \varepsilon_n^2$.

5 - une droite qui minimise $\varepsilon_1'^2 + \ldots + \varepsilon_n'^2$.

6 - une droite qui minimise $\varepsilon_1''^2 + \ldots + \varepsilon_n''^2$.

etc.

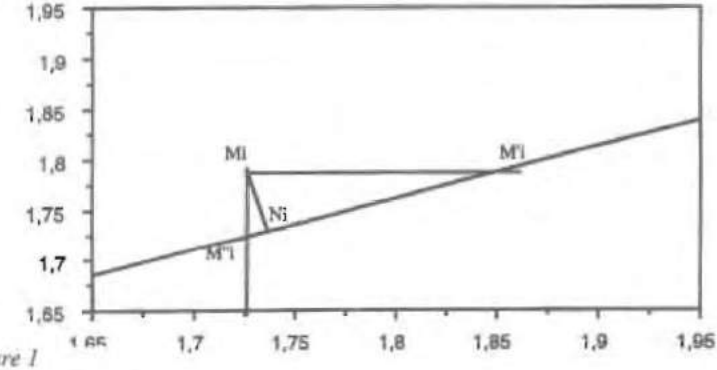


figure 1

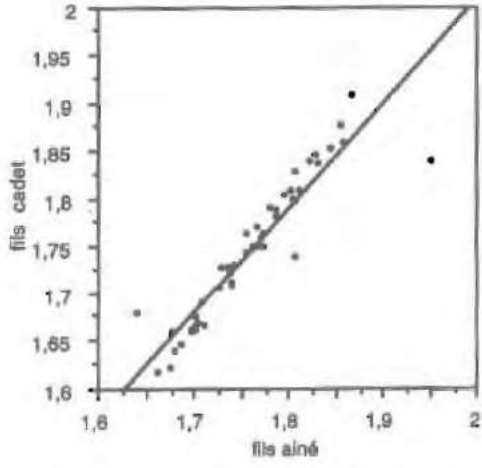
Les critères 1, 2, 3 peuvent être trouvés par des élèves. Les critères 4, 5, 6 sont moins naturels pour eux, et si on les emploie, il convient de les justifier.

Le choix entre les critères 1, 2, 3 se fait en fonction de l'objectif poursuivi : si on veut approximer y_i, $i = 1...n$, par une fonction affine des x_i, alors ε_i'' est l'erreur commise pour le cas n°i et c'est le critère 3 qu'il convient de choisir. Si on veut approximer x_i, $i = 1...n$, par une fonction affine des y_i, alors ε'_i est l'erreur commise pour le cas n°i et c'est le critère 2 qu'il convient de choisir. Si on cherche à ce que les x_i et les y_i jouent un rôle symétrique, alors c'est le premier critère qu'on choisit.

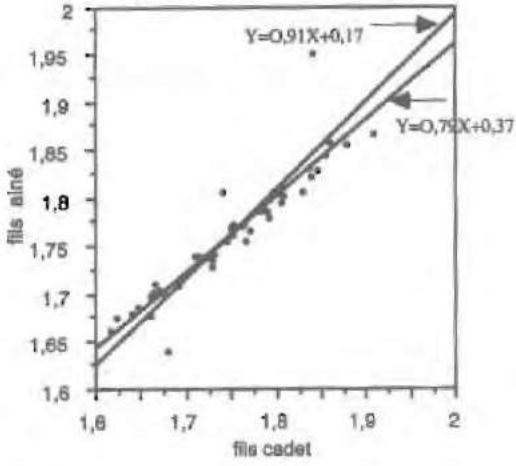
Mais il se trouve que pour les critères 1, 2, 3, on ne peut pas trouver de formules exactes donnant la pente et l'ordonnée à l'origine d'une meilleure droite (il se peut même qu'une telle droite ne soit pas unique). Par contre, pour les critères 4, 5, 6, il y a unicité de la solution et on a des formules simples donnant la droite cherchée. C'est une des raisons majeures de l'utilisation des critères 4,5,6 de préférence aux critères 1, 2, 3.

Le critère 6 fournit la droite d'ajustement linéaire des y_i sur les x_i et le critère 5 la droite d'ajustement linéaire des x_i sur les y_i (si les écarts-type des séries des x_i et de y_i ne sont pas égaux, les pentes de ces deux droites ne sont pas inverses l'une de l'autre cf. Figure 2-).

Le critère 4 fournit "la droite d'ajustement orthogonal", peu utilisée lors



Droite des moindres carrés
$y = 1{,}1x - 0{,}19$



Droite des moindres carrés
$y = 0{,}79x + 0{,}37$

Figure 2

On étudie la taille du frère aîné et du frère cadet dans les familles où il y a au moins deux fils (les tailles sont celles qu'ils ont à 20 ans). On a ainsi un échantillon (ici, il s'agit d'un échantillon simulé) de taille 50 de couples de taille du frère aîné et du frère cadet (x_i est la taille du frère aîné de la famille n°i et y_i est la taille du frère cadet de la même famille).

En haut, droite D d'ajustement linéaire de la taille du fils cadet sur celle du fils aîné. En bas, droite D' d'ajustement linéaire de la taille du fils aîné sur celle du cadet; on a aussi représenté la droite D précédente sur ce graphique (c'est la droite $y = (x + 0{,}19)/1{,}1 = 0{,}91x + 0{,}17$).

qu'on étudie deux variables mais très utilisé lorsqu'on fait de la statistique descriptive multivariée à plus de deux variables.

2.3. Les critères 5 et 6 sont les plus utilisés à cause des modèles de régression linéaire que nous allons brièvement décrire à travers l'exemple des données de la figure 2. On étudie la taille du frère aîné et du frère cadet dans les familles où il y a au moins deux fils (les tailles sont celles qu'ils ont à 20 ans). La taille (notée y) du frère cadet n'est évidemment pas une fonction déterministe de la taille du frère aîné (notée x) et pour une taille donnée du frère aîné, on a une certaine distribution de la taille du frère cadet. On utilise fréquemment pour ce type de données un modèle de régression linéaire dont une hypothèse fondamentale est que la moyenne, notée $\overline{y}_x$, des valeurs d'une quantité y, calculée pour une valeur fixée de x, est une fonction linéaire de x, ce que nous écrirons $\overline{y}_x = \mathbf{a}x + \mathbf{b}$.

Ainsi, dans l'exemple considéré, la taille moyenne du frère cadet pour une taille donnée du frère aîné sera une fonction linéaire de celle-ci.

Les nombres **a** et **b** sont des paramètres théoriques du modèle dont on estimera des valeurs sur un échantillon. Lorsqu'on dispose d'un échantillon (x_i, y_i), $i = 1 \ldots n$, on peut montrer que de bonnes estimations de ces valeurs sont respectivement les coefficients **a*** et **b*** de la droite d'ajustement linéaire des y_i sur les x_i. Dans ce cadre là, la droite $y = \mathbf{a}^*x + \mathbf{b}^*$, qui donne, sur tout un intervalle de valeurs de x, une approximation de la valeur moyenne de y pour une valeur fixée de x, s'appelle droite de régression linéaire de y sur x.

En d'autres termes, la droite d'ajustement linéaire, considérée comme outil descriptif, permet d'approximer les ordonnées des points d'un nuage et ne fait intervenir aucune notion de modèle probabiliste. Cette droite intervient dans la régression linéaire qui, elle, repose sur un modèle probabiliste, et s'appelle alors la droite de régression linéaire.

Références :

- Pierre Dagnélie , *Statistique théorique et appliquée*. Tomes 1 et 2, Pierre Dagnélie - Éditions Les presses agronomiques de Gembloux, 1992 .
-Thomas H. Wonnacott et Ronald J. Wonnacott, *Statistique. économie, gestion, sciences, médecine*. Éditions Economica. 1990.
-C. Robert, *L'empereur et la girafe. Leçons élémentaires de statistique*. Éditions Diderot. 1995.