Un exemple d'utilisation en statistique de la distance d'un point à un plan

Catherine Dufossé

Le nouveau programme de spécialité de la classe de TES comporte en commentaire de « *Exemples de calculs de la distance d'un point à un plan* » la phrase suivante : « *On pourra à cette occasion interpréter géométriquement des cas très simples d'optimisation de fonctions de deux variables sous contrainte en minimisant la distance d'un point à un plan.* »

Cette phrase a plongé dans l'embarras pas mal de gens à Marseille et sans doute ailleurs. Une solution possible du problème nous a été apporté lors d'une réunion de formateurs par Robert Roland, directeur de l'IREM de Marseille. Voici sa forme.

1 - Le problème

L'optimisation à réaliser est le problème de la droite de régression d'une série double.

On s'occupe d'une série double, $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$, i variant de 1 à n. Elle est habituellement représentée par un nuage de points $B_i(x_i, y_i)$.

Le problème est de trouver une droite approximant " au mieux " le nuage de points. La façon usuelle de préciser ce mieux consiste à chercher la droite D telle que la somme $\sum_{i=1}^{n} B_i M_i^2$ soit minimum, où M_i est le point de

la droite D d'abscisse x_i. Nous appellerons par la suite m_i l'ordonnée du point M_i.

L'étude de ce problème est au programme de la classe de TES : pratique du calcul dans la partie obligatoire, démonstration des formules dans la spécialité. La façon de traiter ce problème exposée ici n'est pas la méthode usuelle, qui consiste à trouver le minimum d'une fonction polynôme de degré deux dépendant d'un paramètre. Elle utilise la distance d'un point à un plan.

2- L'idée

On peut utiliser la géométrie de l'espace du programme de spécialité pour résoudre ce problème à condition de se contenter de trois points. On se place alors dans l'espace muni d'un repère orthonormé, et un point représente alors une série. Ainsi, les séries (x_1, x_2, x_3), (y_1, y_2, y_3), (m_1, m_2, m_3) deviennent les coordonnées de trois points : X, Y, M. La somme à minimiser, $\sum_{i=1}^{n} B_i M_i^2$ $= \sum_{i=1}^{n} (y_i - m_i)^2$, n'est autre que le carré de la distance YM.

Or la condition d'alignement des points M_i se traduit par l'appartenance du point M à un plan. Pour les élèves, on peut faire se calcul au cas par cas, mais il est très joli de constater que les équations $(m_i = ax_i + b)$, pour $i = 1, 2, 3$, traduisent d'une part que les points $M_i(x_i, y_i)$ appartiennent dans le plan à une droite d'équation (y=ax+b) et d'autre part que le point M appartient dans l'espace au plan de repère $(O, \vec{u}, \overrightarrow{OX})$, où $\vec{u}$ est le vecteur de coordonnées (1,1,1), puisqu'elle peut se lire aussi : $\overrightarrow{OM} = a\overrightarrow{OX} + b\vec{u}$.

Le minimum de $\sum_{i=1}^{n} B_i M_i^2 = \sum_{i=1}^{n} (y_i - m_i)^2$ est donc atteint lorsque M est le projeté orthogonal de Y sur ce plan. La solution est unique, et le minimum de la somme $\sum_{i=1}^{n} B_i M_i^2$ est le carré de la distance du point Y à ce plan.

3- En classe

Il est bien sûr restrictif de n'utiliser que trois points, mais on expliquera aux élèves que le calcul se fait pour n points en dimension n et ils peuvent comprendre qu'on simplifie le problème pour le mettre à leur portée.

C'est un lien intéressant entre deux parties du programme : il apporte une justification de l'étude de la géométrie de l'espace dans cette série en proposant une vision différente d'un problème qui semblait purement calculatoire, et il prépare les élèves de spécialité à des développements ultérieurs.

Sur le plan des idées, ce changement de point de vue, très classique en mathématiques (pensons à l'espace dual d'un espace vectoriel), est une nouveauté intéressante pour eux, tout en restant tout à fait abordable.

On constatera enfin que le calcul de la somme des résidus dans ce cas ne manque pas d'élégance.

Exemple de calcul :

Les points B_i sont : $B_1(3,1)$, $B_2(6,2)$, $B_3(9,2)$.

Les points M_i sont : $M_1(3,m_1)$, $M_2(6,m_2)$, $M_3(9,m_3)$.

La condition d'alignement des points M_i s'écrit :

$$3(m_3 - m_1) - 6(m_2 - m_1) = 0\,,$$

ce qui traduit l'appartenance du point M au plan P de l'espace d'équation 3 (Z-X)-6 (Y-X)=0, ou encore : X-2Y+Z=0.

Remarque : cette équation était prévisible *a priori* dans ce cas particulier, car elle traduit le fait que le point M_2 est le milieu du segment $[M_1M_3]$ puis[illegible]t 9.

[illegible]ns [illegible]nce où on a t[illegible]porté le problème, on cherche la distance mi[illegible]mum [illegible]n point du [illegible] P et le point Y de coordonnées (1,2,2). On sait que ce [illegible]um est at[illegible] lorsque le point du plan est le projeté orthogonal du po[illegible]ur le pla[illegible]

Un vecte[illegible]mal à [illegible] le vecteur de coordonnées (1,-2,1). Une équation param[illegible] de [illegible]erpendiculaire à P passant par Y est donc :

$$\begin{cases} x = 1 + t \\ y = 2 - 2t \\ z = 2 + t \end{cases}$$

Le para[illegible]rnissant les coordonnées du point d'intersection de cette droite[illegible]an P est la solution de l'équation du premier degré :

$(1+t)-2(2-2t)+(2+t)=0$. On obtient : $t=\frac{1}{6}$. Le projeté M de Y sur le plan P a donc pour coordonnées : $\begin{cases} m_1=\frac{7}{6} \\ m_2=\frac{10}{6} \\ m_3=\frac{13}{6} \end{cases}$.

Les points M_i de la droite d'ajustement ont donc pour coordonnées : $M_1\left(3,\frac{7}{6}\right), M_2\left(6,\frac{5}{3}\right), M_3\left(9,\frac{13}{6}\right)$. L'équation de la droite de régression se calcule alors comme l'équation d'une droite passant par deux de ces points. Elle s'écrit : $y=\frac{x}{6}+\frac{2}{3}$.

Quant à la somme des résidus $\sum_{i=1}^{n} B_i M_i^{\,2}$, elle est égale à YM^2, c'est-à-dire au carré de la distance de Y au plan P, qui est égale à :

$$YM=\frac{|1-4+2|}{\sqrt{1+4+1}}=\frac{1}{\sqrt{6}}.$$

La somme des résidus est donc égale à $\frac{1}{6}$.

Le calcul usuel de ces résultats fournit bien sûr les mêmes valeurs. Il est plutôt plus long si on le fait à la main. Il est certainement intéressant de le faire faire aussi aux élèves pour qu'ils aient sur le même problème les deux points de vue.

4- Et ce n'est qu'un début…

J'ai découvert par la suite, en particulier en épluchant les documents issus des réunions inter-académiques (je me souviens d'un document sur cette question en provenance de Toulouse) que ce point de vue sur la statistique permettait une vision géométrique de tout le programme de TES en statistique. C'est trop joli pour que je le passe ici sous silence.

Quelques pistes, donc, que les intéressés développeront, et toujours avec trois points seulement pour rester en dimension trois :

Si l'on associe, à la série $(x_i)_{i=1}^3$ représentée par le point X, le « point moyen » $\overline{X}$ de coordonnées $(\overline{x},\overline{x},\overline{x})$ on s'aperçoit que ce point n'est autre que le projeté orthogonal de X sur la droite Δ d'équation (x=y=z). Il en résulte que les points représentant des séries de même moyenne sont les points d'un plan perpendiculaire à cette droite.

L'écart-type de cette série est tout simplement le tiers de la distance $X\overline{X}$, distance du point X à la droite Δ. Il en résulte que les points représentant des séries de même écart-type sont les points d'un cylindre d'axe Δ. C'est ainsi que les séries de même écart-type et de même moyenne sont représentées par les points d'un cercle centré sur Δ et situé dans un plan perpendiculaire à Δ.

On peut aussi répondre aisément à la question suivante : deux réels ne peuvent pas avoir une somme quelconque quand leur produit est donné. Une série peut-elle avoir une moyenne quelconque lorsque son écart-type est fixé (ou vice-versa) ? Cette représentation géométrique du problème prouve que oui sans calcul : tout plan perpendiculaire à Δ coupe n'importe quel cylindre d'axe Δ.

Examinons ensuite le coefficient de corrélation linéaire entre les séries $(x_i)_{i=1}^3$ et $(y_i)_{i=1}^3$. Il n'est rien d'autre que le cosinus de l'angle des vecteurs $\overrightarrow{X\overline{X}}$ et $\overrightarrow{Y\overline{Y}}$: sa formule le montre d'elle-même. Voilà une explication sans aucun calcul du fait qu'il est compris entre -1 et 1.

Les séries $(x_i)_{i=1}^3$ et $(y_i)_{i=1}^3$ sont bien corrélées si ces deux vecteurs $\overrightarrow{X\overline{X}}$ et $\overrightarrow{Y\overline{Y}}$, sont « presque » colinéaires. Voilà une explication des mystérieuses « valeurs limites » pour décider si l'approximation linéaire est ou non valide. Vous trouverez dans des manuels différents des valeurs différentes : supérieur à 0.7, à 0.8, à 0.86. Le problème, en fait, est de savoir si cet angle est proche de zéro ou de pi, et chacun donne sa propre interprétation de cette « proximité ».

J'avoue que depuis que j'ai découvert cette vision des choses, les statistiques à deux variables ont pris pour moi… du relief !