STATISTIQUE ET GEOMETRIE

Dans cet article, nous allons essayer d'établir des liens entre les statistiques d'une part, et la géométrie d'autre part. En effet, on lit souvent que des statisticiens célèbres ont montré (ou ont eu l'intuition) de résultats de statistique par des considérations de géométrie ; de plus, certains d'entre nous doivent ou devront enseigner ce qu'il est convenu d'appeler "l'analyse des données" (modules D4* en BTSA).
En restant modeste, essayons de voir ce qu'il en est pour les paramètres les plus classiques (moyenne, variance...).
Certains résultats seront généralisés aux variables aléatoires réelles, ces résultats seront donnés en italiques.

1. QUELQUES NOTATIONS :

Pour la partie statistique :

On considère une variable statistique, notée X dans la suite. On suppose que l'on dispose de n observations (n valeurs) pour X, notées $x_1, x_2, ..., x_n$. On note respectivement $\overline{x}$ et s^2 la moyenne et la variance de la série $(x_1, x_2, ..., x_n)$.
En fait, ici nous ne travaillerons pas avec la variable X, mais seulement avec la série des n observations $(x_1, x_2, ..., x_n)$ de cette variable.

Pour la partie géométrie :

On considère l'espace vectoriel euclidien muni du produit scalaire usuel, c'est-à-dire si $\vec{u}(a_1, a_2, ..., a_n)$ et $\vec{v}(b_1, b_2, ..., b_n)$ sont deux vecteurs de cet espace vectoriel alors on a : $\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} a_i b_i$ et $\left\|\vec{u}\right\|^2 = \sum_{i=1}^{n} a_i^2$.

On note $\vec{1}$ le vecteur de coordonnées (1, 1, ..., 1).
De même, on considère l'espace affine euclidien $\mathbb{R}^n$ associé à cet espace vectoriel dont l'origine est notée O.

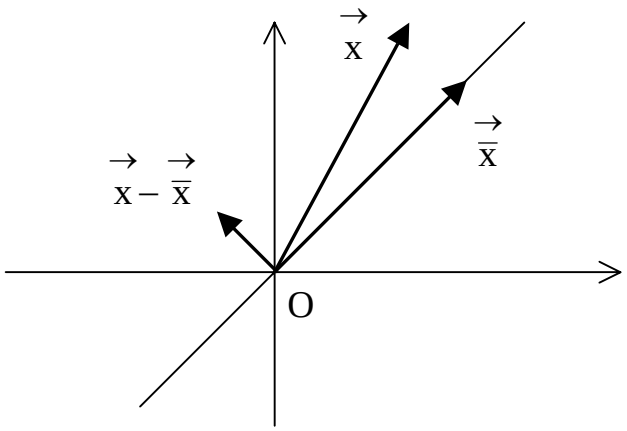
2. A PROPOS DE LA MOYENNE :

Considérons le vecteur $\vec{x}(x_1, x_2, ..., x_n)$, ce vecteur peut être considéré comme le vecteur « image » de la distribution statistique $(x_1, x_2, ..., x_n)$.

Par définition, le vecteur $\vec{\bar{x}}(\bar{x}, \bar{x}, ..., \bar{x})$ est le vecteur « moyenne ».

Par exemple, dans $\mathbb{R}^2$:

Considérons le vecteur $\vec{x}(2\,;4)$. On définit alors les vecteurs $\vec{\bar{x}}(3\,;3)$ et $\vec{x}-\vec{\bar{x}}(-1\,;1)$.



Résultat 1 :

$$\vec{x}\,.\,\vec{1}=\sum_{i=1}^{n}x_i \qquad \text{soit} \qquad \vec{x}\,.\,\vec{1}=n\bar{x}$$

L'application qui à tout vecteur $\vec{u}$ fait correspondre le réel $\vec{u}\,.\,\vec{1}$ est une application linéaire (en fait une forme linéaire).

L'opération, qui à tout n-uplet $(x_1, x_2, ..., x_n)$ associe sa moyenne $\bar{x}$, peut donc être considérée comme le produit scalaire $\vec{x}\,.\,\vec{1}$ (à la constante multiplicative $\frac{1}{n}$ près).

Cette opération a donc les mêmes propriétés que le produit scalaire. *On retrouve ainsi la propriété de « linéarité » de l'espérance mathématique, c'est-à-dire, en termes de variables aléatoires :*

Si a et b sont deux nombres réels alors E(aX + b) = aE(X) + b.

Les différentes moyennes pondérées peuvent aussi être vues sous la forme précédente (voir annexe 1).

On déduit aussi du résultat 1 que les vecteurs $\vec{x}$ et $\vec{1}$ sont orthogonaux si et seulement si $\bar{x}$ = 0.

Résultat 2 :

$$\left(\vec{x}-\vec{\bar{x}}\right).\vec{1}=\sum_{i=1}^{n}(x_i-\bar{x})$$

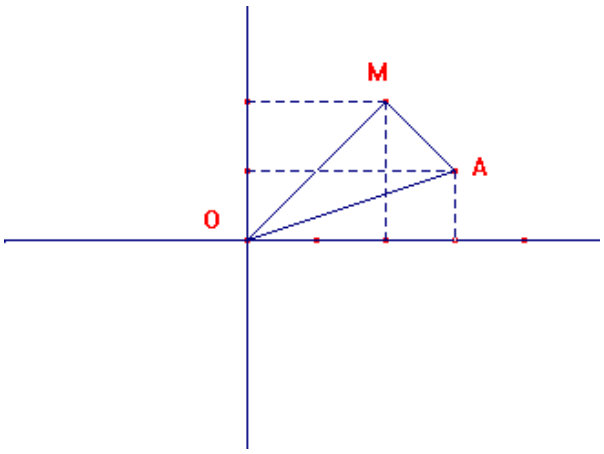
$\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=0$. On en déduit $\left(\vec{x}-\vec{\bar{x}}\right).\vec{1}=0$.

Les vecteurs $\vec{x}-\vec{\bar{x}}$ et $\vec{1}$ sont donc orthogonaux.

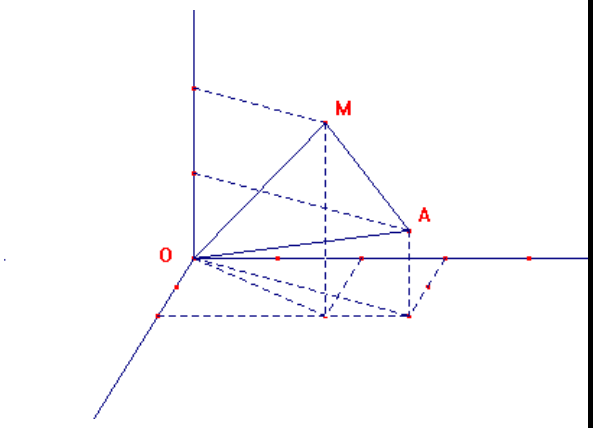
Par suite, les vecteurs $\vec{x}-\vec{\bar{x}}$ et $\vec{\bar{x}}$ sont orthogonaux (car $\vec{\bar{x}}=\bar{x}.\vec{1}$).

Le vecteur $\vec{\bar{x}}$ est le projeté orthogonal du vecteur $\vec{x}$ sur le sous-espace $\mathbb{R}\vec{1}$, c'est-à-dire la droite vectorielle engendrée par le vecteur $\vec{1}$.
En termes savants, c'est le théorème de la projection orthogonale.

Quelques représentations graphiques :



A est le point de coordonnées (3;1).
Le vecteur $\vec{x}$ est le vecteur $\overrightarrow{OA}$.
Le point M est le point de coordonnées (2;2).
Le vecteur $\vec{\bar{x}}$ est le vecteur $\overrightarrow{OM}$.
Le vecteur $\vec{x}$ - $\vec{\bar{x}}$ est le vecteur $\overrightarrow{MA}$.



A est le point de coordonnées (3;2;1).
Le vecteur $\vec{x}$ est le vecteur $\overrightarrow{OA}$.
Le point M est le point de coordonnées (2;2;2).
Le vecteur $\vec{\bar{x}}$ est le vecteur $\overrightarrow{OM}$.
Le vecteur $\vec{x}$ - $\vec{\bar{x}}$ est le vecteur $\overrightarrow{MA}$.

3. A PROPOS DE LA VARIANCE :

Résultat 3 :

$$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$ c'est-à-dire $$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = ns^2$$

La variance est donc égale au carré d'une norme divisé par n. Elle en possède donc les mêmes propriétés.

Par exemple, en termes de variables aléatoires :

Si a et b sont deux nombres réels alors $V(aX + b) = a^2V(X)$.

Au passage, on peut remarquer que $\left\| \vec{x} - \vec{\bar{x}} \right\|^2$ est la somme des carrés des écarts à la moyenne, elle est souvent notée SCE (notamment dans le cadre de l'analyse de la variance ou de la régression).

<u>Résultat 4</u> :

$$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \vec{x}^2 - \vec{\bar{x}}.\vec{x}$$

En effet : $\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \left(\vec{x} - \vec{\bar{x}} \right).\left(\vec{x} - \vec{\bar{x}} \right)$

$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \left(\vec{x} - \vec{\bar{x}} \right).\vec{x}$ car les vecteurs $\vec{x} - \vec{\bar{x}}$ et $\vec{\bar{x}}$ sont orthogonaux.

D'où, $\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \vec{x}^2 - \vec{\bar{x}}.\vec{x}$

Il en résulte : $s^2 = \frac{1}{n}\left(\vec{x}^2 - \vec{\bar{x}}.\vec{x} \right)$ c'est-à-dire $s^2 = \frac{1}{n}\left(\vec{x}^2 - \bar{x}.\vec{1}.\vec{x} \right)$

ou encore $s^2 = \frac{1}{n}\vec{x}^2 - \bar{x}^2$

On retrouve la formule de KŒNIG-HUYGENS, c'est-à-dire, *en termes de variables aléatoires :*

$$V(X) = E(X^2) - [E(X)]^2$$

En remarquant que l'égalité $\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \vec{x}^2 - \vec{\bar{x}}.\vec{x}$ peut s'écrire

$$\left\|\vec{x}-\vec{\bar{x}}\right\|^2=\left\|\vec{x}\right\|^2-\left\|\vec{\bar{x}}\right\|^2 \quad \text{ou encore} \quad \left\|\vec{x}\right\|^2=\left\|\vec{\bar{x}}\right\|^2+\left\|\vec{x}-\vec{\bar{x}}\right\|^2$$

on constate que la formule de KŒNIG-HUYGENS n'est rien d'autre que le théorème de PYTHAGORE exprimé dans $\mathbb{R}^n$.

Plaçons nous dans l'espace affine $\mathbb{R}^n$ rapporté à un repère orthonormal $(O, \vec{e_1}, \vec{e_2}, ..., \vec{e_n})$.
Étant donné un point $M(x_1, x_2,...,x_n)$ de $\mathbb{R}^n$, on note, pour tout i de {1, 2, ..., n}, $A_i(x_i, x_i, \ldots,x_i)$ le projeté orthogonal de M sur le sous-espace $\mathbb{R}\vec{e_i}$.
L'isobarycentre G des points A_i a pour coordonnées $(\bar{x}, \bar{x}, \ldots, \bar{x})$.

En appliquant la formule de LEIBNIZ (voir annexe 2),

$$\sum_{i=1}^{n}\alpha_i MA_i^2=\sum_{i=1}^{n}\alpha_i MG^2+\sum_{i=1}^{n}\alpha_i GA_i^2$$

avec M = O et pour tout i de {1, 2, ..., n} $\alpha_i=\alpha$ où α est un nombre réel non nul, on obtient :

$\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}\bar{x}^2+\sum_{i=1}^{n}(\bar{x}-x_i)^2$, c'est-à-dire $\left\|\vec{x}\right\|^2=\left\|\vec{\bar{x}}\right\|^2+\left\|\vec{x}-\vec{\bar{x}}\right\|^2$.

On retrouve donc la formule de KŒNIG-HUYGENS.

Ce résultat reste vrai pour une moyenne pondérée.

4. <u>A PROPOS DE LA SOMME DES CARRÉS DES ÉCARTS</u> :

La formule de LEIBNIZ permet de retrouver le résultat de statistique suivant :

La moyenne $\bar{x}$ est l'unique réel qui minimise la somme des carrés des écarts à un réel *a*.
Soit, exprimé autrement, $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \inf_{a\in\mathbb{R}}\left(\sum_{i=1}^{n}(x_i-a)^2\right)$.

En effet, dans le cas général :

$\sum_{i=1}^{n} \alpha_i MA_i^2$ est minimum si et seulement si $\sum_{i=1}^{n} \alpha_i MG^2 = 0$

Les α_i sont tous strictement positifs, donc

$\sum_{i=1}^{n} \alpha_i MA_i^2$ est minimum si et seulement si M = G.

Remarquons aussi que si on appelle inertie du système de points pondérés (A_i, α_i) par rapport à un point M de $\mathbb{R}^n$, le nombre réel, noté I(M), défini par I(M) = $\sum_{i=1}^{n} \alpha_i MA_i^2$ alors l'inertie est minimum si et seulement si M = G.
Soit en termes de physiciens, « c'est par rapport au centre de gravité du système que les masses ont le plus petit moment d'inertie ».

Pour vous persuader de l'intérêt de la géométrie en statistique, vous trouverez en annexe 3 une preuve du résultat suivant :

Pour une série statistique donnée (x_i), l'écart-type est supérieur ou égal à l'écart absolu moyen : $\sigma \geq \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$.

5. <u>UNE PREMIÈRE CONCLUSION</u> : (en attendant une éventuelle suite)

Suite à cette modeste introduction, on peut se poser quelques questions pour aller plus loin, en voici trois.

a) Dans ce qui précède, nous avons considéré un paramètre de tendance centrale, la moyenne, et un paramètre de dispersion, l'écart-type.
L'idée géométrique sous-jacente étant : trouver le nombre réel *a* tel que la distance euclidienne classique entre le point $M(x_1, x_2, ..., x_n)$ et le point A(*a*, *a*, ..., *a*) soit minimale.

Plus généralement, on peut se poser la question suivante :

> Comment résumer une série statistique par un paramètre de tendance centrale et un paramètre de dispersion ?

Cette question peut aussi s'énoncer sous la forme :

Soit d une distance associée à $\mathbb{R}^n$, déterminer le réel *a* tel que la distance entre le point $M(x_1, x_2, ..., x_n)$ et le point A(*a*, *a*, ..., *a*) soit minimale.

Et dans ce cas, la valeur *a* ainsi déterminée est le paramètre de tendance centrale associé à la distance et d(A,M) caractérise la dispersion.

Par exemple, pour la distance euclidienne classique, on obtient la moyenne et l'écart-type (l'écart-type est égal à d(A,M) divisé par $\sqrt{n}$).

On voit bien que si l'on prend une autre distance, on aura a priori d'autres paramètres ; est ce que le réel *a* sera unique ?
Pour une série statistique donnée, quelle(s) distance(s) choisir ? Pourquoi choisir telle distance plutôt qu'une autre ?

b) Pour une distance donnée associée à $\mathbb{R}^n$, l'idée de minimiser ou maximiser l'inertie (par rapport à ...) est une des idées directrices de l'analyse des données.

c) Enfin,

Qu'en est-il dans le cas de deux ou de plusieurs variables ?
Comment interpréter géométriquement la covariance ?
En quoi le calcul matriciel permet de poser les problèmes sous une autre forme et de simplifier des preuves ?
Faut-il travailler dans l'espace des variables ou dans l'espace des individus ?
...

Bref, il reste encore beaucoup de questions à élucider ...

<u>ANNEXES</u> :

1°) Soit $(\alpha_1, \alpha_2, ..., \alpha_n)$ n réels <u>positifs</u>, on appelle moyenne pondérée de la série $(x_1, x_2, ..., x_n)$ le nombre réel $\frac{1}{\sum \alpha_i} \sum_{i=1}^{n} \alpha_i x_i$.

Les résultats énoncés précédemment pour la moyenne arithmétique restent vrais, il suffit de considérer (à la place du vecteur $\vec{1}$) le vecteur $\vec{p}(\alpha_1,\alpha_2,...,\alpha_n)$, (p comme poids).

2°) Soit $(A_i ; \alpha_i)$, $i \in \{1, 2, ..., n\}$ un système de points pondérés, on appelle fonction scalaire de LEIBNIZ la fonction qui à tout point M de l'espace associe le nombre réel $\sum_{i=1}^{n} \alpha_i MA_i^2$. Le résultat que nous utilisons est le suivant :

Si G désigne le barycentre de ce système de points, alors on a pour tout point M de l'espace, on a $\sum_{i=1}^{n} \alpha_i MA_i^2 = \sum_{i=1}^{n} \alpha_i MG^2 + \sum_{i=1}^{n} \alpha_i GA_i^2$.

3°) L'inégalité de CAUCHY-SCHWARZ s'écrit $|\vec{u}.\vec{v}| \leq \|\vec{u}\| \times \|\vec{v}\|$ soit encore

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2} .$$

D'où, en prenant $a_i = \frac{1}{\sqrt{n}} |x_i|$ et $b_i = \frac{1}{\sqrt{n}}$, on a $\frac{1}{n} \sum_{i=1}^{n} |x_i| \leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$ (1)

Considérons une série statistique $(x_1, x_2, ..., x_n)$. Quitte à faire un changement de variable, on peut supposer que la moyenne est nulle. L'inégalité (1) signifie que l'écart absolu moyen est inférieur ou égal à l'écart-type.

- = - = - = - = - = - = - = -