Statistiques à deux variables(corrélation et causalité)

I – Première approche :

Lorsqu'il existe un lien entre deux caractères d'une population, on peut définir une série statistique à deux variables. Nous allons étudier un exemple.

1) Tableau de données :

La résistance thermique (en m^2.°C / W) d'un mur dépend de l'épaisseur (en cm) d'un isolant.

Epaisseur	2	4	6	8	10	12	15	20
Résistance	0,83	1,34	1,63	2,29	2,44	2,93	4,06	4,48

2) Nuage de points :

On peut placer les points correspondant dans un repère orthogonal. On obtient un nuage de points. Il semble intéressant de trouver une fonction f telle que la courbe y=f(x) soit le plus proche possible des points du nuage. C'est le problème de l'ajustement.
(si c'est une droite, on parle d'ajustement affine, mais pas toujours – ici il y a corrélation, mais ce n'est pas toujours le cas)

3) Point moyen :

Dans un premier temps, on peut calculer le point dont les coordonnées sont les moyennes des coordonnées des points du nuage.
Ici : G=

II - Ajustement affine :

1) Méthode graphique : on place la règle laissant moitié des points de part et d'autre. Il convient de prendre une droite qui passe par le point moyen. Il reste à noter deux points sur cette droite et en trouver l'équation.

2) Méthode de Mayer : on coupe le nuage de points en deux et on calcule le point moyen de chaque partie (la droite passe toujours par le point moyen).
Ici, $G_1=$ et $G_2=$

3) Droite de régression : soit D d'équation $y=ax+b$ une droite d'ajustement. $M_i(x_i;y_i)$ un point du nuage. P_i est le point d'abscisse sur D.
On appelle droite de régression de y en x la droite D telle que $\sum_{i=1}^{n} M_i P_i^2$ soit minimale.
On défini de même la droite de régression de x en y en remplaçant l'abscisse par l'ordonnée dans le choix des P_i.

Cette méthode s'appelle la méthode des moindres carrés.

4) Covariance : $\text{cov}(x;y)=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

5) Equations des droites de régression :
La droite de régression D de y en x a pour équation y= ax + b avec $a=\frac{cov(x;y)}{V(x)}$ et où b vérifie $\bar{y}=a\bar{x}+b$ avec V(x) la variance de la série à une variable x.

6) <u>Coefficient de corrélation linéaire</u> : $r = \frac{cov(x\,;y)}{\sigma(x)\sigma(y)}$. On admet que $-1 \leq r \leq 1$.

si $r^2 = 1$, C'est un ajustement parfait.
si r^2 est proche de 1, il y a une bonne corrélation entre x et y.
(la notion de proche de 1 dépend du sujet étudié – 0,5 dans certains secteur du bâtiment – 0,999 dans l'industrie)

7) Utilisation de la calculatrice pour l'exemple étudié :

III – <u>Changements de variables</u> :

Dans les cas où l'ajustement affine n'est pas justifié, il est parfois possible de résoudre le problème par un changement de variables qui permet d'obtenir un nouveau nuage de points pour lequel un ajustement affine convient.

Exemple du livre p264.