STATISTIQUES À DEUX VARIABLES

Table des matières

I. Statistiques à une variable (rappels)

Définition : Soient $x_{1,} y_{1,} \ldots , x_p$ les valeurs distinctes d'une série statistique et $n_{1,} n_{2,} \ldots , n_p$ les effectifs correspondants.

Moyenne : $\bar{x}=\dfrac{n_1 x_1+n_2 x_2+\ldots+n_p x_p}{n_1+n_2+\ldots+n_p}$. **Variance** : $V=\dfrac{n_1 x_1^2+n_2 x_2^2+\ldots+n_p x_p^2}{n_1+n_2+\ldots+n_p}-\bar{x}^2$.

Écart-type : $\sigma=\sqrt{V}$.

Dans la pratique, on préfère l'écart type à la variance car l'écart type peut être comparé à l'ordre de grandeur des valeurs, ce qui n'est pas le cas de la variance.

Exemple : les 31 élèves d'une classe de Première ont obtenu les notes suivantes à un contrôle de mathématiques.

Notes : x_i	7	8	9	10	11	12	13	14
Effectif : n_i	1	5	4	12	5	3	0	1

Moyenne : $\bar{x}=$ ....

....

....

Variance : $V=$ ....

....

....

Écart-type : $\sigma=$ ....

En pratique, pour l'écart-type, on prend la calculatrice (voir *fiche calculatrice*) et on trouve :

$\sigma \approx$ ....

II. Statistiques à deux variables

II.1 Késako ?

Dans certains cas, il semble exister un lien entre deux caractères d'une ***série statistique à deux variables****, par exemples : entre le poids et la taille d'un nouveau-né, entre la consommation et la vitesse d'une voiture, etc. Ce lien n'est pas nécessairement une relation de cause à effet : la vente des crèmes solaires semble liée à celle des crèmes glacées sans qu'aucune des deux soit la cause ou la conséquence de l'autre (toutes deux sont certainement des conséquences d'un autre phénomène : l'ensoleillement).*

Dans ces cas là, il peut être intéressant d'étudier simultanément deux caractères d'une même population. Les résultats peuvent alors être présentés sous différentes formes (tableaux, graphiques, etc).

Exemple 1

Au cours du premier trimestre de cette année, une entreprise a lancé la commercialisation d'un accessoire « C » nécessaire à la pose de son produit « B ». On dispose des quantités vendues par zones de vente :

Zones	*Nombre d'unités de B vendues :* x_i	*Nombre d'unités de C vendues :* y_i
1	4 000	2 400
2	2 000	1 200
3	6 000	3 000
4	3 000	1 500
5	3 000	1 200
6	6 000	2 700

Exemple 2

Pour des véhicules légers de la gammes de 9-11 CV fiscaux, roulant en palier (ou en descente), on a relevé les consommations moyennes et les vitesses suivantes :

Vitesse en km/h : x_i	10	20	30	40	50	60	70	80	90
Consommation en l/100 km : y_i	16,5	11,5	9,0	7,5	6,8	6,6	7,0	7,5	9,0

> **Définition** : Sur des individus d'une population, on réalisé simultanément N observations de 2 caractères quantitatifs x et y.
> L'ensemble des N couples $(x_1;y_1),\dots,(x_N;y_N)$ où x_1 et y_1, … , x_N et y_N sont les valeurs observées de x et de y, est appelée ***série statistique à 2 variables* x *et* y**.

Le plan étant muni d'un repère, nous pouvons associer au couple $(x_i;y_i)$ de la série statistique double, le point M_i de coordonnées x_i et y_i. L'ensemble des points M_i obtenus constitue le ***nuage de points*** représentant la série statistique.

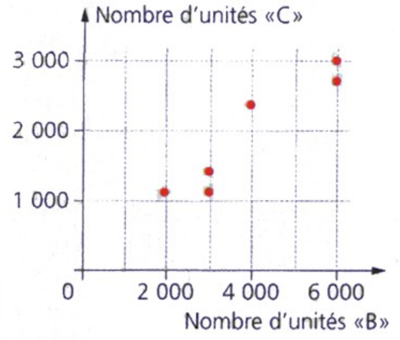


Figure 1

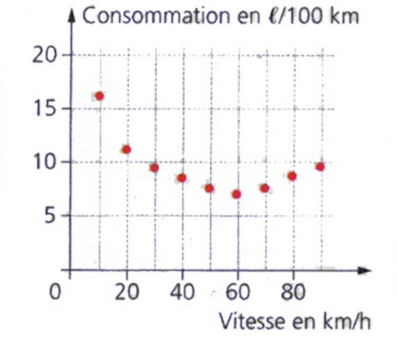


Figure 2

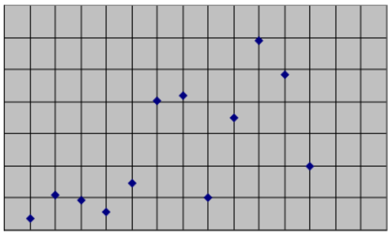
II.2 Ajustement affine

Dans les exemples 1 et 2, nous obtenons les nuages des figures 1 et 2. Le nuage étant dessiné, on peut essayer de trouver une fonction f telle que la courbe d'équation $y=f(x)$ « passe le plus près possible » des points du nuage. C'est ce qu'on appelle un ***problème d'ajustement***.

▪ Dans l'exemple 1, on peut penser qu'en première approximation, une droite D peut être tracée au voisinage de ces 6 points. On dit alors que l'on a un ***ajustement affine***.

▪ Dans l'exemple 2, un ajustement affine ne semble pas approprié : on peut penser à « approcher » le nuage par une parabole.

On peut trouver des nuages dont les points sont dispersés de façon quelconque, notamment lorsqu'il n'existe aucun lien entre x_i et y_i, par exemple :



Lorsqu'on pense pouvoir réaliser un ajustement affine d'un nuage, il peut sembler intéressant, avant de tracer la droite, de placer le point dont l'abscisse est la moyenne $\bar{x}$ des abscisses x_i, et l'ordonnée la moyenne $\bar{y}$ des ordonnées y_i.

Définition : on appelle ***point moyen*** d'un nuage de n points M_i de coordonnées $(x_i ; y_i)$ le point G de coordonnées $x_G = \bar{x}$ et $y_G = \bar{y}$.

Dans l'exemple 1, le point moyen G est (4 000 ; 2 000) car :

....

....

....

....

....

....

....

....

[*placer ce point sur la figure*]

II.2.1. Méthode graphique

On considère à nouveau le nuage de points de l'exemple 1.
On se propose, à partir des quantités vendues, de faire des prévisions de vente de produit C pour d'autres chiffres de ventes du produit B. Un moyen d'y parvenir est de tracer « au jugé » une droite D passant le plus près possible des points du nuage et d'admettre que les chiffres de vente y_i du produit C et x_i du produit B sont liés par l'équation $y = ax + b$ de D.

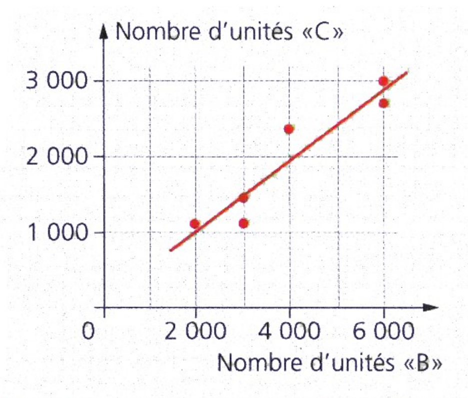


Figure 3

Remarque : la méthode graphique a l'avantage de sa simplicité apparente et de sa rapidité ; cependant, chaque utilisateur de cette méthode peut tracer une droite différente, ce qui pose le problème du choix entre plusieurs propositions.

II.2.2. **Méthode des moindres carrés**

Une entreprise s'intéresse au lien entre ses dépenses publicitaires et son chiffre d'affaires : elle recueille les données suivantes, exprimées en millions d'euros, portant sur cinq périodes où les dépenses publicitaires sont notées d_1, d_2, …, d_5 et les chiffres d'affaires c_1, c_2, …, c_5.

Dépenses publicitaires : d_i	0,5	2,0	2,9	4,5	5,6
Chiffre d'affaires : c_i	35	37	75	92	90

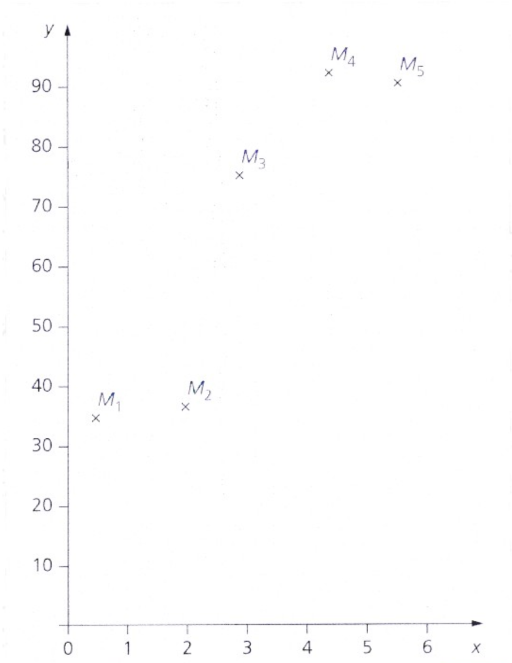


Figure 4

On représente ces données par cinq points M_i dans un repère où les dépenses publicitaires sont en abscisse et les chiffres d'affaires en ordonnée : $x_i = d_i$ et $y_i = c_i$.

Ce nuage de cinq points semble suffisamment allongé pour justifier un ajustement affine...

Mais le problème est de déterminer quelle droite est susceptible de remplacer « au mieux » ce nuage de points !

La ***méthode des moindres carrés*** consiste à déterminer la droite d'ajustement $\mathscr{D}$ telle que la somme $P_1M_1^2 + \ldots + P_5M_5^2$ soit minimale.

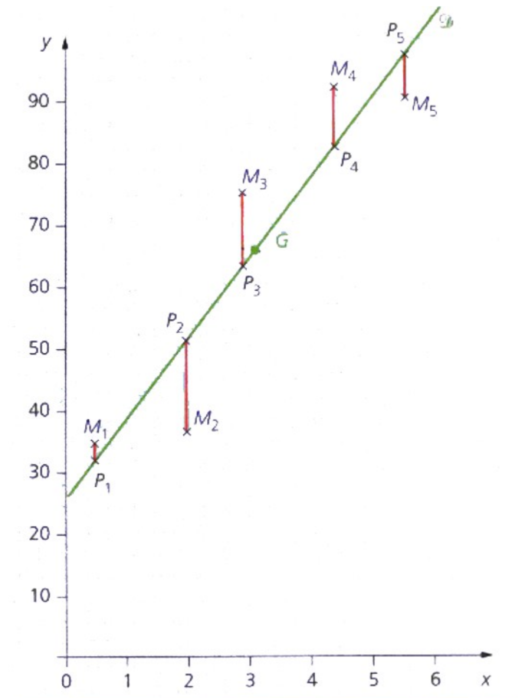
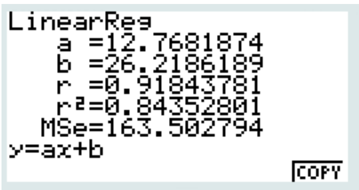


Figure 5

Cette droite s'appelle ***droite de régression linéaire*[1] *de y en x***, et passe par le point moyen du nuage (propriété admise).

La calculatrice[2] (ou un tableur) nous donne directement une équation de cette droite :



Ici, on considère donc que la droite $\mathscr{D}$ a pour équation réduite : $y = 12{,}768\,x + 26{,}219$.

Cette relation nous permet d'**estimer**, par exemple, le montant du chiffre d'affaires c (en millions d'euros) associé à des dépenses publicitaires d de 3,5 millions d'euros :
$c = 12{,}768 \times 3{,}5 + 26{,}219$ donc $c \approx 70{,}907$.

Remarque : graphiquement, on peut observer sur la figure 5 que le point de la droite D d'abscisse 3,5 a une ordonnée proche de 70,9.

1 En latin, *gradus* signifie « pas » ou « marche ». *Régression* signifiait donc à l'origine « marcher en arrière ». Le statisticien anglais Francis Galton, cousin de Charles Darwin, introduisit ce terme en 1885. Travaillant sur l'hérédité, il cherchait à « expliquer » la taille des fils en fonction de celle de leur père : il constata que lorsque le père était plus grand que la moyenne, son fils avait tendance à être plus petit que lui et, a contrario, que lorsque le père était plus petit que la moyenne, son fils avait tendance à être plus grand que lui. Il y avait donc régression au sens courant du terme... Ce travail amena Galton a développer sa théorie *regression toward mediocrity*.

2 <u>**CASIO**</u> : page 331 <u>**TEXAS INSTRUMENTS**</u> : page 327