Statistiques à une variable (rappels)

Définition : Soient $x_1, y_1, \dots, x_p$ les valeurs distinctes d'une série statistique et $n_1, n_2, \dots, n_p$ les effectifs correspondants.

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

Écart-type : $\sigma = \sqrt{V}$.

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

Exemple 1 : les 31 élèves d'une classe de Première ont obtenu les notes suivantes a un contrôle de mathématiques.

Notes x_i	7	8	9	10	11	12	13	14
Effectifs n_i	1	5	4	12	5	3	2	1

Moyenne : Variance : Écart-type :

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

Statistiques à deux variables

Présentation

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 2
Au cours du premier trimestre de cette année, une entreprise a lance la commercialisation d'un accessoire "C" nécessaire a 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 3
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.

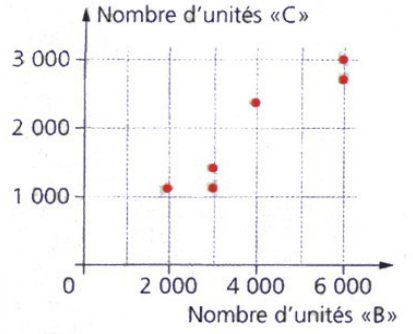


Figure 1

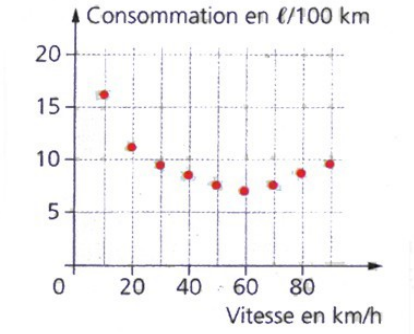


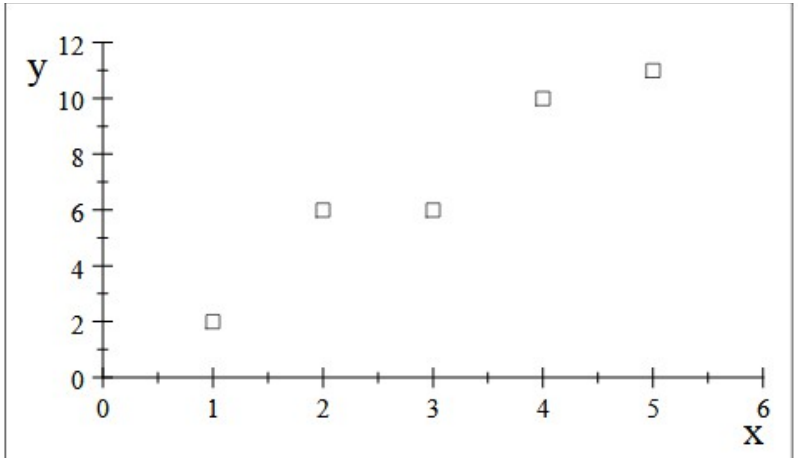
Figure 2

Droite d'ajustement affine

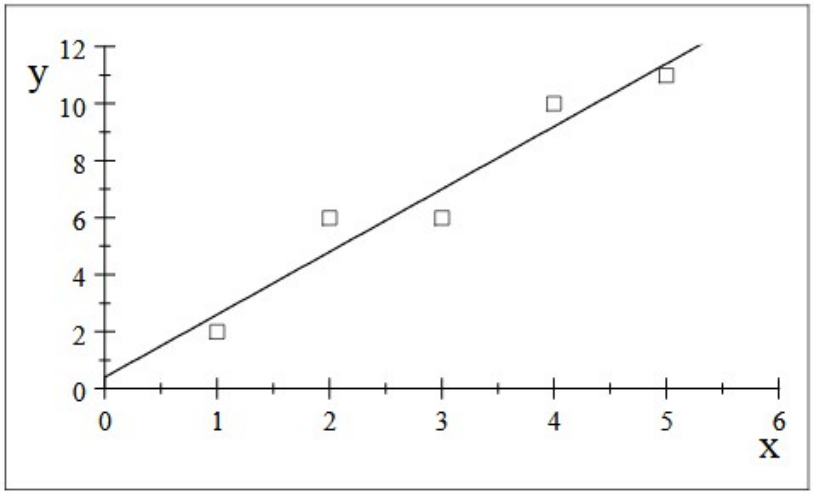
Méthode graphique

x_i	2	3	5	1	4
y_i	6	6	11	2	10

Exemple 4 : on considère la série statistique à 2 variables ci-contre
On obtient le graphique en « Nuage de points » ci-dessous



L'objectif réside alors dans la construction d'une droite (d) passant « au plus près » de tous les points M_i du Nuage



Méthode de MAYER

x_i	2	3	5	1	4
y_i	6	6	11	2	10

Exemple 4 :
on considère la série statistique à 2 variables ci-contre

L'objectif est de diviser le Nuage de points en 2 Nuages distincts

- Le Nuage formé des 3 premiers points M_1, M_2, M_3
- Le Nuage formé des 2 derniers points M_4 et M_5

On calcule alors les 2 points moyens des 2 Nuages de points

1er point moyen $G_1(x_1 ; y_1)$:

$x_1 = \frac{2+3+5}{3} = 3{,}33$ et $y_1 = \frac{6+6+11}{3} = 7{,}67$ donc $G_1(3{,}33 ; 7{,}67)$

2ème point moyen $G_2(x_2 ; y_2)$:

$x_2 = \frac{1+4}{2} = 2{,}5$ et $y_2 = \frac{2+10}{2} = 6$ donc $G_2(2{,}5 ; 6)$

L'équation de la droite de MAYER passant par G_1 et G_2 est $(d) : y = a\,x + b$

$a = \frac{y_2 - y_1}{x_2 - x_1} = \frac{6 - 7{,}67}{2{,}5 - 3{,}33} = 2{,}012$

$b = y_1 - a \times x_1 = 7{,}67 - 2{,}012 \times 3{,}33 = 0{,}97$

donc la droite de MAYER est $(d) : y = 2{,}012\,x + 0{,}97$

Méthode des Moindres carrés

x_i	2	3	5	1	4
y_i	6	6	11	2	10

Exemple 4 :
on considère la série statistique à 2 variables ci-contre

On calcule le point moyen du Nuage de points (complet) $G(\overline{x} ; \overline{y})$

avec $\overline{x} = \frac{2+3+5+1+4}{5} = 3$ et $\overline{y} = \frac{6+6+11+2+10}{5} = 7$ donc $G(3 ; 7)$

On calcule la Variance de x : $Var(x)=\frac{\sum x_i^2}{n}-(\overline{x})^2$

on obtient $Var(x)=\frac{2^2+3^2+5^2+1^2+4^2}{5}-(3)^2=\frac{55}{5}-9=2$

On en déduit l'écart-type de x : $\sigma(x)=\sqrt{Var(x)}=\sqrt{2}\simeq 1{,}41$

On calcule la covariance du couple $(x\,;y)$: $Cov(x,y)=\frac{\sum x.y}{n}-(\overline{x}).(\overline{y})$

on obtient $Cov(x,y)=\frac{2\times 6+3\times 6+5\times 11+1\times 2+4\times 10}{5}-3\times 7=\frac{127}{5}-35=4{,}4$

L'équation de la droite (Δ) des « moindres carrés » est :

$(\Delta): y=ax+b$ avec $a=\frac{Cov(x,y)}{Var(x)}$ et $b=\overline{y}-a\times\overline{x}$

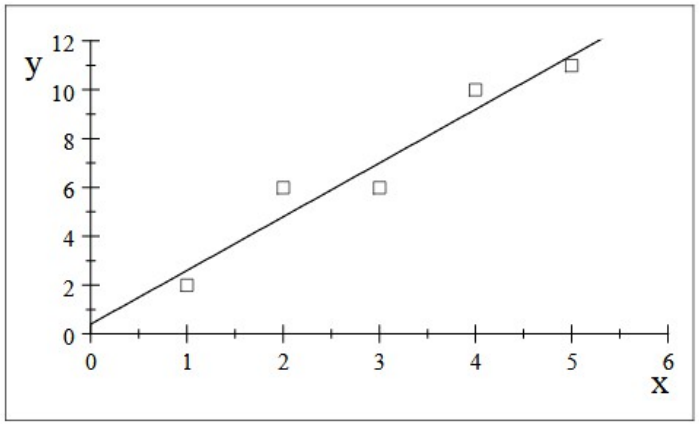
on obtient ici : $a=\frac{4{,}4}{2}=2{,}2$ et $b=7-2{,}2\times 3=0{,}4$ donc $(\Delta): y=2{,}2x+0{,}4$

on obtient le graphique ci-dessous

Cette droite est aussi appelée **« droite de régression »** de y en x

Remarque : En latin, *gradus* signifie "pas" ou "marche". *Régression* signifiait donc a l'origine "marcher en arriere". Le statisticien anglais Francis Galton, cousin de Charles Darwin, introduisit ce terme en 1885.

Travaillant sur l'heredite, il cherchait a expliquer la taille des fils en fonction de celle de leur pere : il constata que lorsque le pere etait plus grand que la moyenne, son fils avait tendance a etre plus petit que lui et, a contrario, que lorsque le pere etait plus petit que la moyenne, son fils avait tendance a etre plus grand que lui. Il y avait donc regression au sens courant du terme... Ce travail amena Galton a developper sa theorie *regression toward mediocrity*



Interpolations & Extrapolations

Interpolation linéaire

x_i	2	3	5	1	4
y_i	6	6	11	2	10

*On reprend les données de l'**Exemple 4***

On souhaite déterminer la valeur de y sachant que $x=3{,}2$
on obtient : $y=2{,}2\times 3{,}2+0{,}4=7{,}44$

On souhaite déterminer la valeur de x sachant que $y=8$

on obtient : $2{,}2x+0{,}4=8$ donc $2{,}2x=7{,}6$ donc $x=\frac{7{,}6}{2{,}2}=3{,}45$

Extrapolation linéaire

x_i	2	3	5	1	4
y_i	6	6	11	2	10

*On reprend les données de l'**Exemple 4***

On souhaite déterminer la valeur de y sachant que $x=7{,}5$
on obtient : $y=2{,}2\times 7{,}5+0{,}4=16{,}9$

On souhaite déterminer la valeur de x sachant que $y=15$

on obtient : $2{,}2x+0{,}4=15$ donc $2{,}2x=14{,}6$ donc $x=\frac{14{,}6}{2{,}2}=6{,}64$

Complément : Coefficient de corrélation

On appelle "coefficient de corrélation" du couple (x,y) le nombre positif : $r=\frac{Cov(x,y)}{\sigma(x).\sigma(y)}$; ce nombre est utilisé afin de comparer plusieurs ajustements

- on déduit que $0\leqslant r\leqslant 1$
- si $0\leqslant r\leqslant 0{,}2$ alors il n'y a aucune corrélation entre x et y
- si $0{,}2\leqslant r\leqslant 0{,}4$ alors il y a une corrélation faible de y en x
- si $0{,}4\leqslant r\leqslant 0{,}6$ alors il y a une corrélation modérée de y en x
- si $0{,}6\leqslant r\leqslant 0{,}8$ alors il y a une corrélation forte de y en x
- si $0{,}8\leqslant r\leqslant 1$ alors il y a une corrélation très forte de y en x