Analyse des séries statistiques à deux variables

1) Introduction:

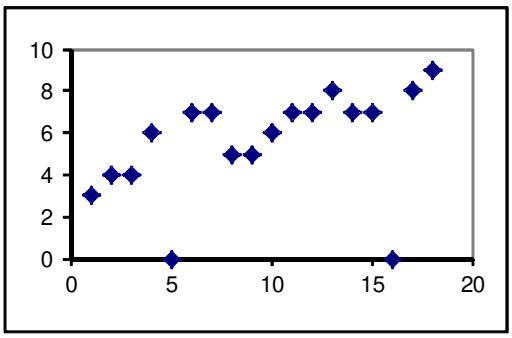
Il arrive fréquemment que l'on observe conjointement deux caractères X et Y dans une même population pour déterminer s'il existe une corrélation entre les deux (âge et taille des enfants entre 0 et 20 ans, prix des terrains à la Roche sur Yon au m² et années).
Exemple :
L'étude porte sur l'influence d'un apport d'aliment concentré sur la croissance de faons au cours de leur premier hiver. Le tableau statistique suivant donne la quantité x de concentré consommé (en grammes) par jour et par animal et la croissance y de l'animal (en grammes) par jour.

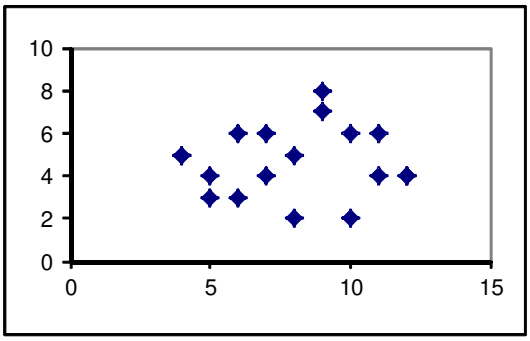
X : quantité de concentré en g	410	420	600	720	750	940	960	1020
Y : gain de poids par jour en g	22	38	40	50	48	76	72	80

La question posée est: Y'-a-t-il un lien entre la prise en poids des faons et l'aliment choisi ?

2) Nuage de points :

On considère une population de N individus et deux caractères quantitatifs représentés par les variables statistiques X et Y.
Chaque individu est représenté par un indice i $(0 \le i \le n)$.
Pour le i-ème individu, la valeur de X est notée x_i et celle de Y est notée y_i.
Dans un repère orthogonal (le plan), le i-ème individu est représenté par le point de coordonnées (X_i , Y_i)
Le nuage de points est l'ensemble des points $M_i(x_i, y_i)$.





Si un point $M_i(x_i, y_i)$ représente plusieurs individus (le couple(X_i , y_i) a un effectif $n_i > 1$) alors le point M(X_i , Y_i) est affecté du coefficient n_i.
Le nuage de points pondérés est l'ensemble des points M(X_i , Y_i) affectés de leurs coefficients n_i.

Si les points semblent être sur une courbe particulière, par exemple une droite (voir courbe), cela peut indiquer qu'il existe une corrélation entre les variables X et Y.

Remarque :

Le problème de corrélation entre deux caractères est étudié dans de nombreux domaines (Commerce, Médecine, Économie, Industrie, ….).

Exemple :

- X peut-être la puissance d'une voiture et Y sa consommation.
- X la consommation de cigarette d'un individu et Y sa durée de vie.

3) Le point Moyen d'un nuage de points ou le barycentre :

On appelle point moyen d'un nuage de points, le point G dont les coordonnées sont respectivement la moyenne des abscisses et la moyenne des ordonnées des points du nuage. Ce point se note $G(\bar{X},\bar{Y})$ avec :

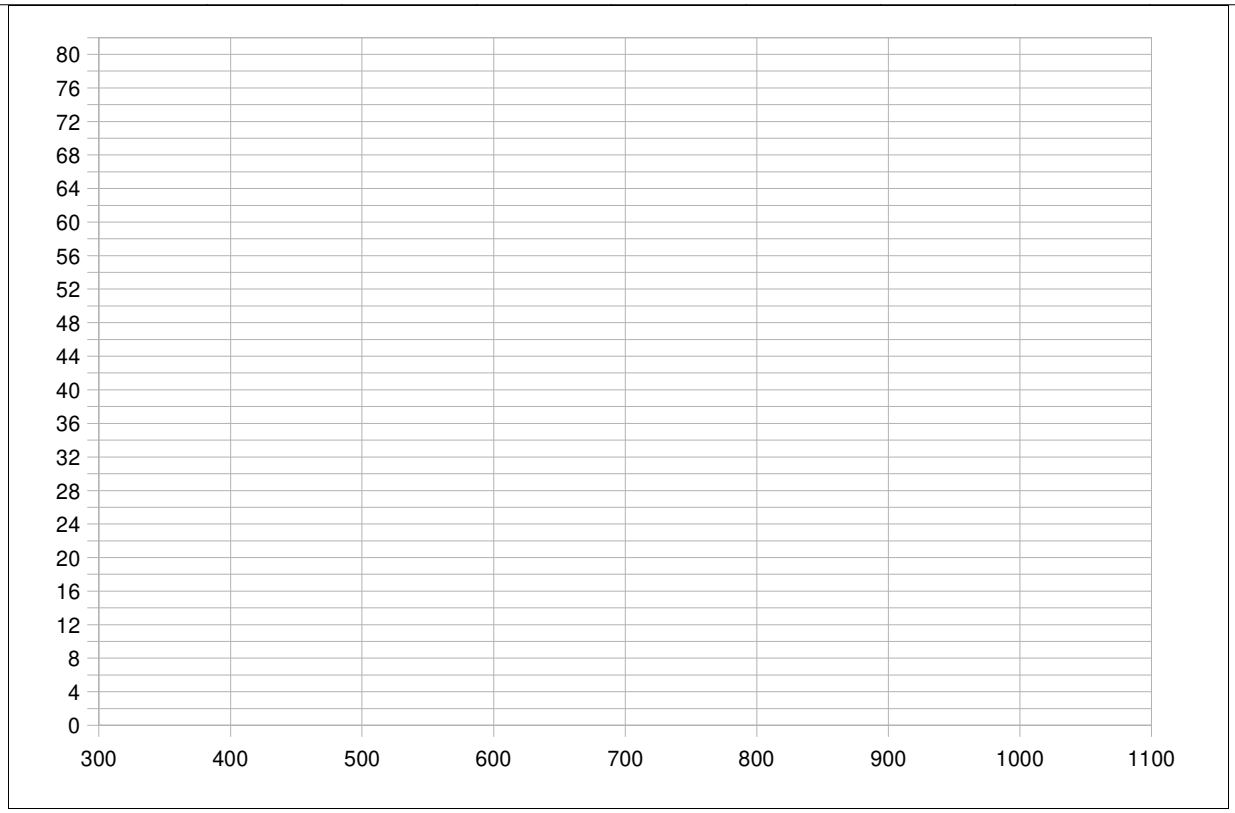
$\overline{X}$ La moyenne de la variable aléatoire X.

$\overline{Y}$ La moyenne de la variable aléatoire Y.

Exemple:

Donner le nuage de point et calculer le point moyen des variables X et Y

X : quantité de concentré en g	410	420	600	720	750	940	960	1020
Y : gain de poids par jour en g	22	38	40	50	48	76	72	80



Le point moyen G est $\bar{x}=$ $\bar{y}=$

4) Covariance

Définition :

la **covariance** est un nombre permettant d'évaluer le sens de variation de deux variables et de qualifier l'indépendance de ces variables

la covariance de X et Y se calcule par :

$$\operatorname{cov}(X,Y)=\sigma_{xy}=\frac{1}{N}\sum_{1}^{N}n_i(x_i-\overline{X})(y_i-\overline{Y})=(\frac{1}{N}\sum_{1}^{N}n_i x_i y_i)-\overline{X}\,\overline{Y}$$

Avec :

$\bar{X}$ La moyenne de la série X. n_i Effectif partiel des séries statistiques X et Y

$\bar{Y}$ La moyenne de la série Y. x_i La i-ème valeur de X. y_i La i-ème valeur de Y.

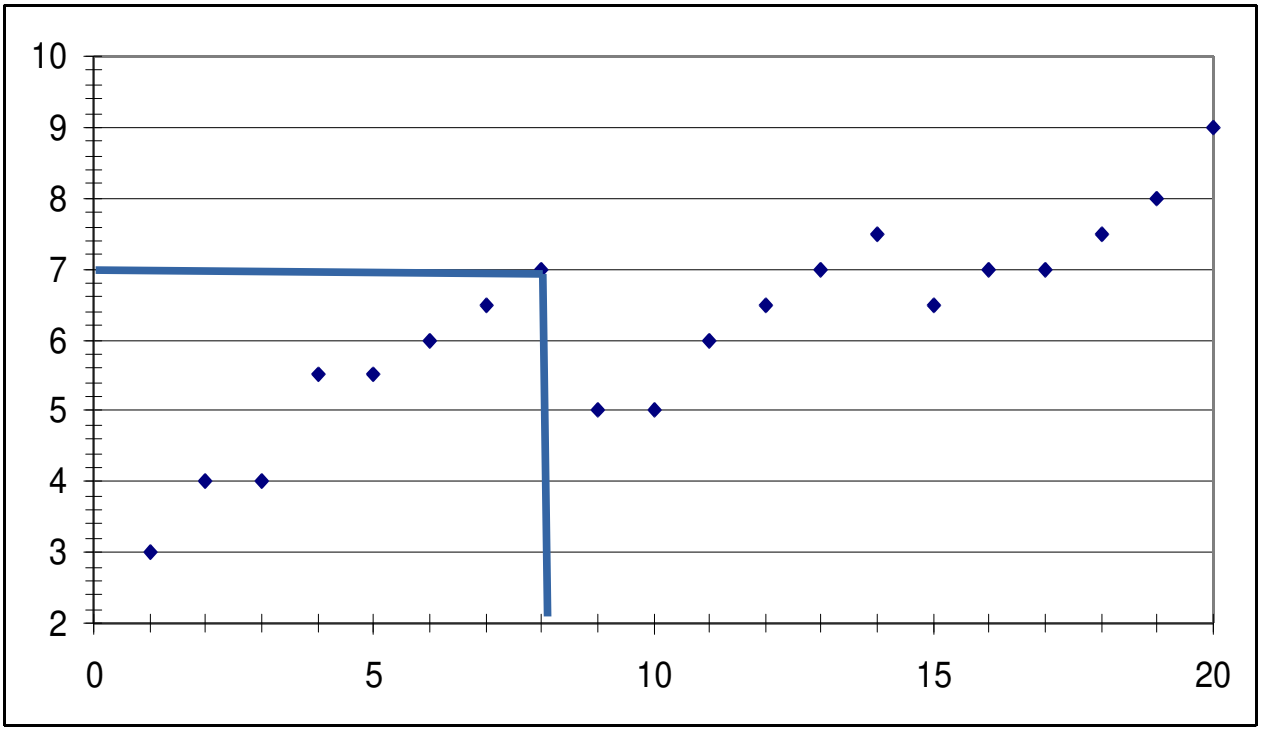
5) Ajustement Linéaire :

La forme du nuage de points pourra suggérer l'existence d'une relation mathématique entre X et Y. La première relation que nous cherchons sera une fonction affine dont la droite est: $y=a\ x+b$

a) Droite de régression

Considérons une série statistique à deux variables représentée par un nuage de points longiligne justifiant un ajustement linéaire.

Soit une droite D quelconque non parallèle à l'axe X'X et non parallèle à l'axe Y'Y sur laquelle tout point M_i se projette en H_i suivant l'axe Y'Y et en Q_i suivant la direction X'X.



b) Méthode des moindres carrés :

La méthode consiste à déterminer une droite (D) qui passe le plus proche possible de tous les points du nuage de points. Pour trouver cette droite il faut rendre les écarts M_iH_i suivant la direction Y'Y ou les écarts M_iQ_i suivant la direction X'X les plus faibles possibles.

i) Ajustement linéaire de y en fonction de x

Définition :

On appelle droite de régression de y en fonction de x la droite (D) obtenue telle que : $S = \sum (M_iH_i)^2$ soit minimale.

La droite (D) ainsi obtenue est appelée droite d'ajustement linéaire de y en fonction x par la méthode des moindres carrés.

Cette droite (D) a pour équation $y = ax + b$

Exemple : a) Donner la droite de régression y en fonction de x de la série statistique double suivante:(Utiliser la calculatrice)

X : quantité de concentré en g	410	420	600	720	750	940	960	1020
Y : gain de poids	22	38	40	50	48	76	72	80

par jour en g								

a=

b=

la droite s'écrit donc:

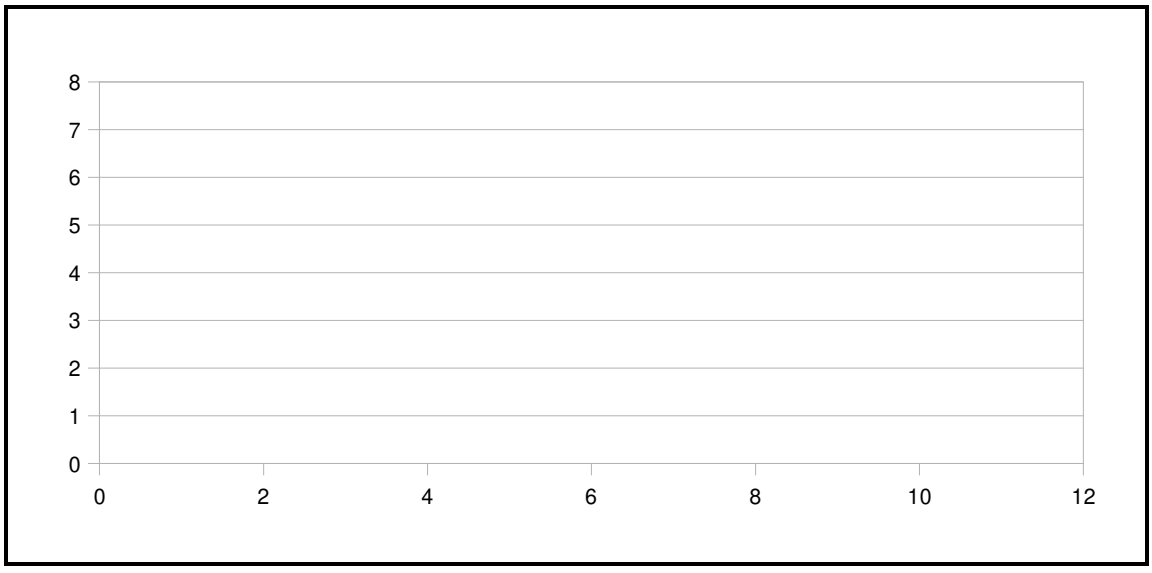
b) Prévoir la prise de poids pour une quantité de concentré de 1030g

Exemple : On considère la série double suivante Notée (X , Y)

xi	1	2	3	4	5	6	7	8	9	10
yi	3	2,5	4	3	5,5	5	4,5	6,5	6	7

1) Donner le nuage de point de la série (X,Y)
2) Calculer $\bar{X}$ et $\bar{Y}$
3) Donner la droite d'ajustement linéaire de y par rapport à x.

Réponse : 1)
Nuage de point



2) le point moyen et les variances :

3) La droite d'ajustement par la méthode des moindres carrés :

6) Coefficient de corrélation ρ :

On appelle coefficient de corrélation le réel ρ défini par : $\rho = \dfrac{\text{cov}(X,Y)}{\sigma_X \ \sigma_Y}$

Ce coefficient est aussi noté ρ (on lit cette valeur directement dans la calculatrice)

Remarque :

$$\rho^2 = a \times a' = \frac{\text{cov}(X,Y)^2}{V(X) \times V(Y)} = \frac{\text{cov}(X,Y)^2}{\sigma_X^2 \times \sigma_Y^2}$$

Rq : Le signe de la pente *a* donne le sens de la corrélation, mais pas sa qualité.

- $a > 0$ corrélation positive
- $a < 0$ corrélation négative
- $a = 0$ pas de corrélation

7) Propriétés du coefficient de corrélation

Le coefficient de corrélation mesure la quantité et la qualité de l'ajustement affine.

a) est toujours $-1 \leq \rho \leq 1$

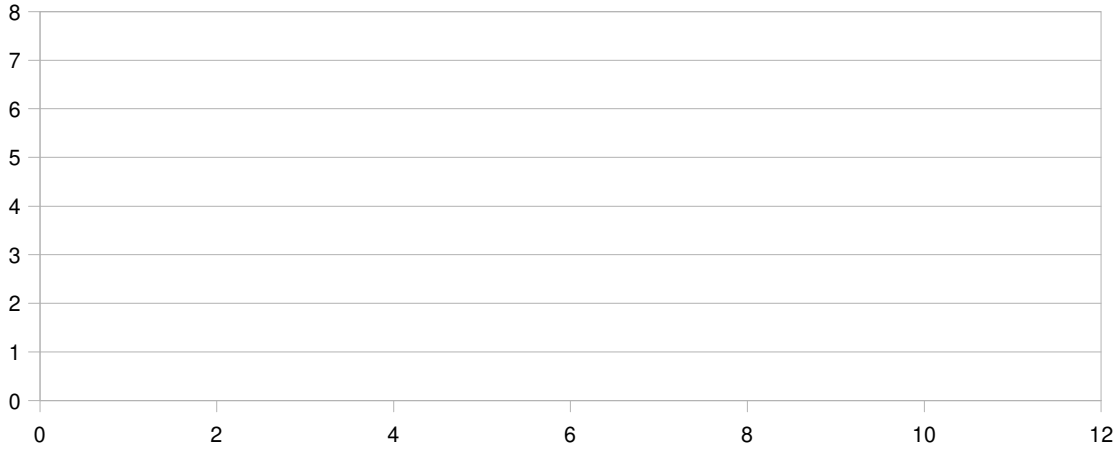
b)Plus il s'éloigne de zéro, meilleure est la corrélation.

- $\rho = +1$ corrélation positive parfaite
- $\rho = -1$ corrélation négative parfaite
- $\rho = 0$ absence totale de corrélation

Exemple : On considère la série double suivante

X	1	2	3	4	5	6	7	8	9	10
Y	4	1,5	4,5	2	3,5	6	4	7	5	7

1) Donner le nuage de point



2) Donner la droite de régression linéaire par la méthode des moindres carrés, le coefficient de régression et commenter l'ajustement

Exercices d' application

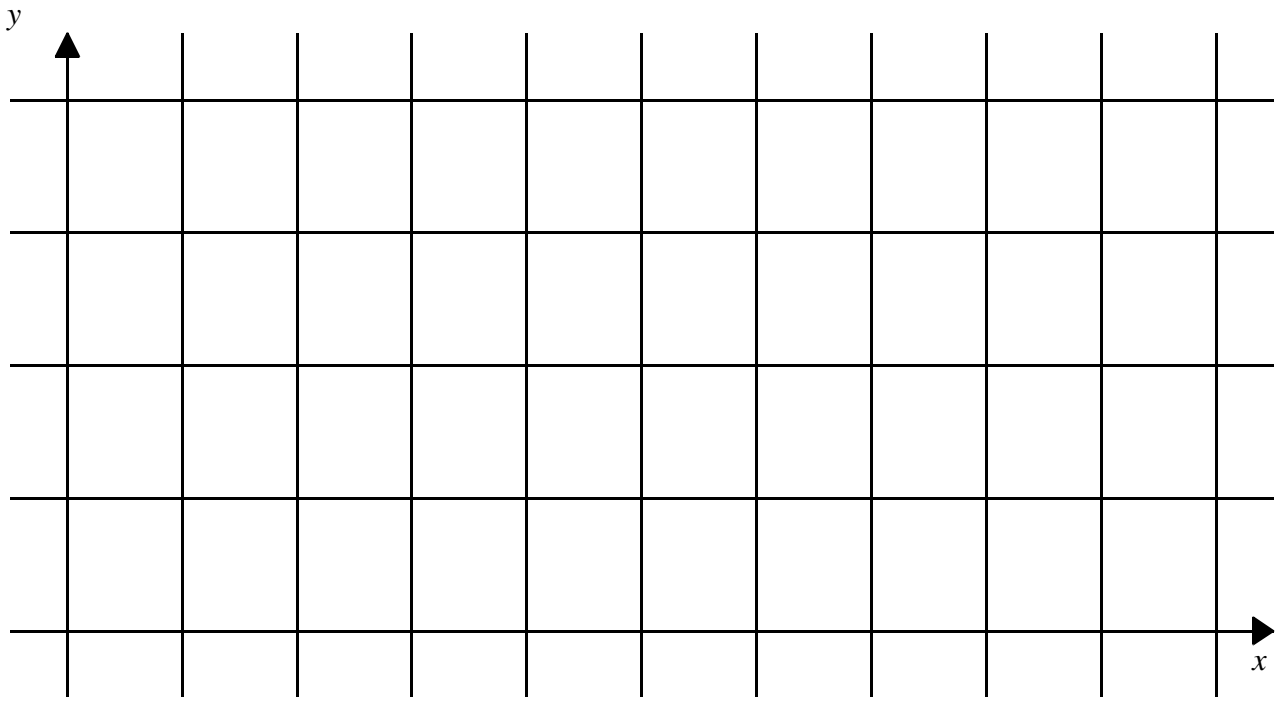
On se propose d'étudier l'évolution du nombre de salariés d'une entreprise depuis sa création. Les résultats des neuf premières années sont consignés dans le tableau suivant , où :

X_i désigne le rang de l'année,

Y_i désigne le nombre de salariés, en dizaines, au premier jour de l'année de rang x_i :

X_i	0	1	2	3	4	5	6	7	8
Y_i	0,1	1	1,6	1,9	2,2	2,4	2,6	2,7	2,8

1) Dans le plan muni d'un repère orthogonal , construire le nuage de points représentant la série double (x_i ; y_i).



2) Préciser la variable expliquée et la variable explicative. Justifier votre réponse

3) On considère le changement de variables $z_i = \exp(y_i)$ où exp désigne la fonction exponentielle de base *e*. *compléter* le tableau des valeurs suivant où on donnera des arrondis des z_i à 10^{-3} près.

X_i	0	1	2	3	4	5	6	7	8
Y_i	0,1	1	1,6	1,9	2,2	2,4	2,6	2,7	2,8
z_i									

4) Déterminer la covariance des variables *X* et *Z*.

$$Cov(X,Z) = \frac{\sum_{1}^{n} x_i y_i}{n} - \bar{x}\bar{z}$$

5) Donner le coefficient de corrélation linéaire des variables *X* et *Z*.

6) Déterminer par la méthode des moindres carrés, une équation de la droite d'ajustement de *Z* en *X*.

7) En déduire une expression de *Y* en fonction de *X*.

8) En utilisant le modèle ainsi défini, évaluer le nombre de salariés de cette entreprise au sixième mois de l'année de rang

Exercice 2

Une entreprise remplit automatiquement des bouteilles de jus de fruits pour les distribuer à la vente en Europe. Cette Entreprise souhaite augmenter ses ventes, pour cela , elle a organisé une enquête sur ses ventes auprès de ses clients pour savoir l'impact du nouveau emballage voir tableau, les x_i représentent l'investissement en millions € et les y_i les recettes en millions €

x_i	30	35	40	45	50
y_i	109	130	148	163	177

1) Construire le nuage de points (x_i , y_i) dans le plan muni d'un repère orthogonal (O ,$\vec{i}$,$\vec{j}$)
(sur la calculatrice)

2) Pour trouver une écriture mathématique des ventes l'entreprise à le choix entre 3 modèles et leurs coefficients de détermination

Modèle numéro	Écriture mathématique	Coefficient de détermination
Modèle 1	$y=-0{,}0486\,x^2+7{,}2657\,x-65{,}086$	$R^2=0{,}9998$
Modèle2	$y=132{,}861\ln(x)-342{,}58$	$R^2=0{,}9999$
Modèle 3	$y=55{,}067\,e^{0{,}0239\,x}$	$R^2=0{,}9755$

a) Quel modèle doit garder cette entreprise ? Justifier votre réponse.

3) On pose pour $1\leq i\leq 5$ $t_i=\ln(3x_i+700)$ où ln désigne le logarithme népérien

a) Compléter le tableau (calcul à 10^{-3} prés)

x_i	30	35	40	45	50
y_i	109	130	148	163	177
t_i					

b) Déterminer une équation de la droite de régression de Y en T par la méthode des moindres carrés, on pourra utiliser les résultats d'une calculatrice 0,01 prés.

c) Donner le coefficient de corrélation linéaire entre T et Y.

d) En déduire une expression de Y en fonction de X.