Statistique à deux variables

1) Introduction

Dans une population on fait deux mesures notés *x* et *y* sur chaque individu statistique. Le but de l'ajustement consiste s'il y a une relation entre ces deux mesures.

En terminale CGRH on recherche s'il existe une relation affine, donc de la forme $y = ax + b$ entre ces deux variables

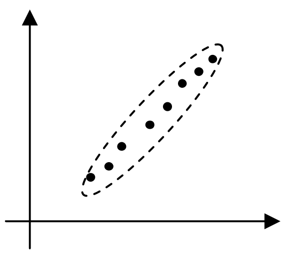
Exemples

On mesure la taille et le poids d'un ensemble de personnes.
On veut vérifier si le poids *y* augmente avec la taille *x*

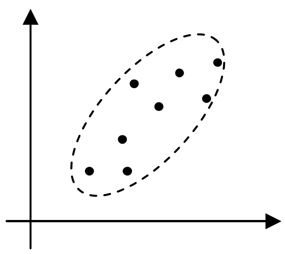
Une entreprise veut étudier l'évolution des son chiffre d'affaire mensuel.
x est le n° du mois et *y* est le chiffre d'affaire

2) Justification d'un ajustement affine

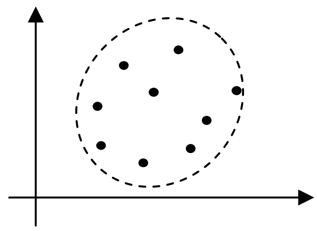
Chaque individu statistique est donc représenté par un couple de nombres (x ;y). On peut donc représenter graphiquement ces couples de nombres par un ensemble de points. Cet ensemble de points est appelé nuage de points. La forme de ce nuage permet de justifier ou non un ajustement affine.



Nuage très allongé
L'ajustement affine est justifié



Nuage allongé
L'ajustement affine reste justifié



Nuage non allongé
L'ajustement affine n'est pas justifié

3) Les deux méthodes d'ajustement affine

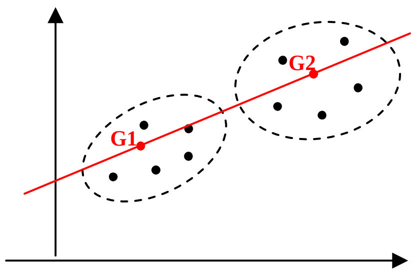
Point moyen

Soient *n* points de coordonnes $(x_1 ; y_1)$; $(x_2 ; y_2)$; $(x_3 ; y_3)$;…………… ; $(x_n ; y_n)$
On appelle point moyen le point $M(x_M ; y_M)$ avec :

$$x_M = \frac{x_1 + x_2 + \dots\dots\dots + x_n}{n}$$

$$y_M = \frac{y_1 + y_2 + \dots\dots\dots + y_n}{n}$$

La méthode de Mayer



On range la série suivant l'ordre croissant de *x*
On considère deux sous nuages égaux ou différents d'une unité
Pour chaque sous nuage on calcule les coordonnées du point moyen. On obtient donc deux points *G1* **et** *G2*
La droite de Mayer est alors la droite *(G1 ; G2)*

La droite de Mayer passe par le point moyen du nuage

La méthode des moindres carrés

Cette méthode de calcul se fait uniquement à la calculatrice.
On obtient trois nombres *a*, *b* et *r*.
a et *b* sont les coefficients de la droite d'ajustement.

La droite des moindres carrés passe par le point moyen du nuage

r est appelé coefficient de corrélation. Ce nombre, compris entre –1 et 1, mesure la qualité de l'ajustement. Si *r* est proche de 1 ou –1 l'ajustement est justifié, Si *r* est proche de 0 l'ajustement n'est pas justifié.
On admet généralement que si :
$0,7 \leq r \leq 1$ l'ajustement affine est croissant et justifié
$-1 \leq r \leq -0,7$ l'ajustement affine est décroissant et justifié.

Comparaison des deux méthodes

La droite de Mayer et la droite des moindres carrés passent toutes deux par le point moyen du nuage de points

Si l'ajustement est justifié ces deux méthodes donnent des résultats très voisins

Si au contraire l'ajustement n'est pas justifié, elles peuvent donner des résultats très différents. Mais dans ce cas un ajustement n'a bien sur aucun intérêt

4) Causalité et corrélation

En statistique on défini une corrélation positive entre deux variables lorsque les deux variables évoluent parallèlement dans le temps, même si elles n'ont rien à voir entre elles.
De même la corrélation sera négative si quand l'une des variables augmente avec le temps l'autre diminue avec le temps

Il y a causalité si les variations de y s'expliquent par celle de la variable x

La causalité implique la corrélation mais la réciproque est fausse : *s'il y a corrélation ceci ne prouve pas la causalité*

Quelques exemples de corrélations sans causalité (exemples tirés de http://www.evoweb.net/stat.htm**)**

Cause cachée :
Dans un article de Science et Avenir sur une étude statistique montrait une corrélation positive entre utilisation de crème solaire et cancer de la peau. Des journalistes pressés en avait conclu un peu vite à la nocivité de la crème solaire

Explication

"*Utilisation de crème solaire*" et "*cancer de la peau*" n'étaient que la conséquence d'une même cause: l'exposition au soleil. Plus on s'*expose au soleil* plus on *risque le cancer de la peau* mais plus aussi on *utilise de crème solaire*

Dans "*Une logique de la communication*", Paul Watzlawick raconte que la plus forte corrélation trouvée dans les années 1950 a été celle entre la consommation de bière sur la côte ouest des USA, et la mortalité infantile au Japon. Cet exemple a été fréquemment repris pour montrer les limites des statistiques et démontrer "*qu'on peut leur faire dire n'importe quoi*". Et en effet beaucoup feront remarquer qu'on ne peut accuser les Américains assoiffés de tuer les Japonais (on remarquera d'ailleurs que personne n'accuse les enfants Japonais d'assoiffer les Américains).

Explication :

Une forte vague de chaleur sur tout le Pacifique, qui avait incité les Américains à augmenter leur consommation de boissons (dont la bière), et qui avait entraîné de graves problèmes sanitaires chez les Japonais pas encore tout à fait remis de la guerre. La corrélation était bien le symptôme d'une causalité.

Aucune cause :
La taille moyenne des japonais a augmenté de 15 cm depuis la fin de la 2ème guerre mondiale alors que la distance entre le Japon et les Etats-Unis augmente de 2 ou 3 cm par an à cause de la dérive des continents. Il y a corrélation parce que les deux phénomènes augmentent avec le temps, mais il n'y a pas bien évidemment la moindre causalité.

5) Quelques exemples d'ajustement affine

Exercice 1 : un exemple d'ajustement justifié

Dans une maternité on a relevé le poids et le taille de 10 nouveaux nés. Les résultats sont consignés dans le tableau suivant :

Enfant	1	2	3	4	5	6	7	8	9	10
Masse en kg	2,5	2,6	2,7	3	3,2	3,3	3,4	3,6	3,8	3,9
Taille en cm	45	46	48	50	51	52	53	54	54	57

On veut savoir si connaissant le poids d'un nouveau né ont peut avoir une idée de sa taille.

1. L'ajustement affine vous parait il justifié ?
2. Faire un ajustement affine de la taille en fonction du poids par la méthode de Mayer:
3. Vérifier que le point moyen est sur la droite d'ajustement
4. Si un bébé pèse 4,2 kg quelle sera sa taille probable ?
5. Refaire les calculs avec la droite des moindres carrés

Exercice 2 : un exemple d'ajustement affine non justifié

Le tableau ci contre montre pour quelques mammifères la durée de gestation en jours et la durée moyenne de vie en années.
Le but de cet exercice est de vérifier si la durée de gestation a une influence sur la durée de vie.

1) représenter graphiquement le nuage de points dans un repère d'unités.
1cm représente 10 jours en abscisses
½ cm représente deux ans en ordonnées
2) Un ajustement affine vous paraît il justifié ?
3) Faites un ajustement affine par la méthode des moindres carrés. Interpréter le coefficient de corrélation r

animal	gestation en jours	durée de vie en années
âne / zèbre	365	12 / 13
babouin	187	20
ours noir	219	18
ours gris	225	25
ours blanc	240	20
castor	105	5
bison	285	15
chameau	406	12
chat	63	12
chimpanzé	230	20
vache	284	15
chevreuil	201	8
éléphant	660	35
élan	250	15
renard	52	7
girafe	457	10
chèvre	151	8
gorille	258	20

Exercice 3 : un ajustement affine justifié et pourtant ce n’est pas le meilleur

On considère le tableau suivant :

x	1	2	3	4	5	6	7	8	9
y	2	5	10	17	26	37	50	65	82

1) Tracer dans un repère le nuage de points.
2) Déterminer par la méthode des moindres carrés une droite d’ajustement affine puis tracer cette droite
3) Interpréter le coefficient de corrélation *r*.
4) Même si le coefficient de corrélation est proche de 1, n’y aurait-il pas une courbe simple qui relie cet ensemble de points ?

Solutions pages suivantes

Exercice 1 :

On pose x le poids en kg et y la taille en cm

1) Les points du nuage sont pratiquement alignés, un ajustement affine est donc bien justifié.

2) Le tableau est dans l'ordre croissant des x on fait donc deux sous nuages, le premier avec les nouveaux nés n° 1 à 5 et le second avec les nouveaux nés n° 6 à 10.

Point moyen du premier sous nuage

$$x_1 = \frac{2,5 + 2,6 + 2,7 + 3 + 3,2}{5} = 2,8$$

$$y_1 = \frac{45 + 46 + 48 + 50 + 51}{5} = 48$$

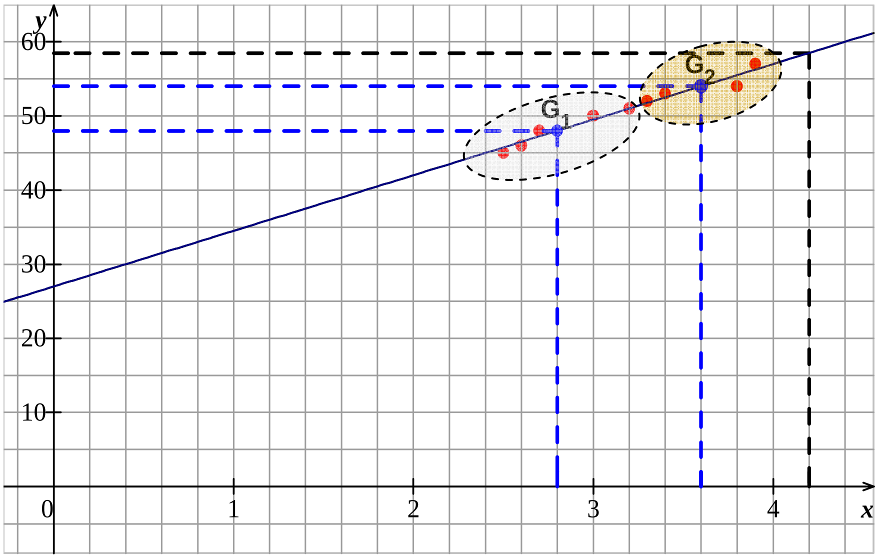
On a donc le point $G_1(2,8\ ;\ 48)$

Point moyen du second sous nuage

$$x_2 = \frac{3,3 + 3,4 + 3,6 + 3,8 + 3,9}{5} = 3,6$$

$$y_2 = \frac{52 + 53 + 54 + 54 + 57}{5} = 54$$

On a donc le point $G_2(3,6\ ;\ 54)$



Calculons l'équation de la droite (G_1 ; G_2)

Son coefficient directeur est :

$$a = \frac{54 - 48}{3,6 - 2,8} = 7,5$$

L'équation de cette droite est y = 7,5x + b

Le point $G_1(2,8\ ;\ 48)$ est un point de cette droite donc :

$48 = 7,5\times2,8 + b \Leftrightarrow 48 - 7,5\times2,8 = b \Leftrightarrow b = 27$

La droite d'ajustement est donc :

y = 7,5x + 27

3) calcul du point moyen *G*

En faisant la moyenne des 10 poids on trouve $x_G = 3,2$

Et la moyenne des 10 tailles donne $y_G = 54$

Le point moyen est donc *G(3,2 ; 54)*

Dans l'équation pour x = 3,2 on a $y = 7,5\times3,2 + 27 = 51 = y_G$

G est donc un point de la droite de Mayer.

4) Si un bébé pèse 4,2 kg alors *x = 4,2*

On remplace x par 4,2 dans l'équation :

$y = 7,5\times4,2 + 27 = 58,8$

Un bébé pesant 4,2 kg mesurera probablement 58 cm

5) Ajustement par les moindres carrés

A la calculatrice on trouve (précision 2 décimales)

a = 7,55 b = 26,85 r = 0,98

La droite des moindre carrés est donc :

y = 7,55x + 26,85

r est très proche de 1, l'ajustement est donc justifié.

Pour x =3,2 on a

$y = 7,55\times3,2 + 26,85 = 51,01$ *(la différence avec 51 s'explique par les arrondis)*

Le point moyen *G* est bien sur la droite des moindres carrés

Pour x = 4,2 on a

$y = 7,55\times4,2 + 26,85 = 58,51$

On retrouve un résultat équivalent à celui obtenu par la méthode de Mayer

Conclusion

Il y a une nette corrélation entre la taille et le poids d'un nouveau né.

Mais il ne s'agit que d'une corrélation l'augmentation du poids ne fait pas grandir un bébé ni l'augmentation de la taille ne le faire grossir.

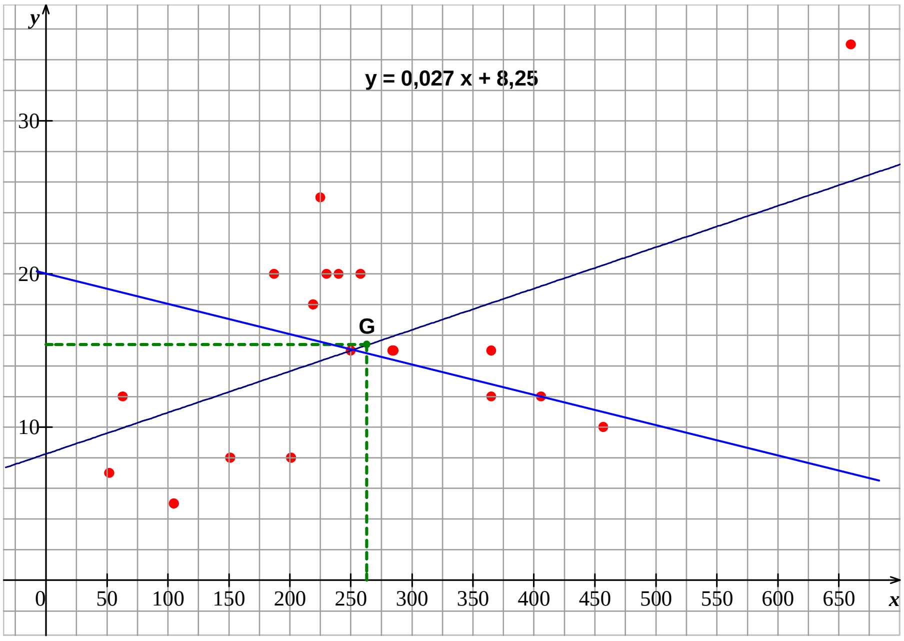
Exercice 2

Le nuage de points est très dispersé, un ajustement affine est donc sans intérêt.
L'ajustement par la méthode des moindres carrés donnerait
a = 0,027 b = 8,25 r = 0,54
Le coefficient de corrélation est inférieur à 0,7
Il n'y a donc aucune corrélation entre la durée de gestation et la durée de vie.

La durée de gestation n'a donc aucune influence sur la durée de vie.

Sur le graphique on a représenté :
la droite des moindres carrés (en noir)
y = 0,027x +8,25
et la droite de Mayer (en bleu)
y = –0,02x + 19,91

Ces deux droites donnent des résultats bien différents ce qui est aussi l'indice qu'il n'y a pas d'ajustement affine



Exercice 3

L'ajustement par la méthode des moindres carrés donnerait
a = 10 b = 17,33 r = 0,97

Un ajustement par la droite des moindres carrés semble donc justifié car *r* est proche de 1.

Cependant on a exactement la relation
$y = x^2 + 1$

Donc la variation de *y* en fonction de *x* n'est pas du tout affine