Statistiques bivariées : corrélation et régression linéaire

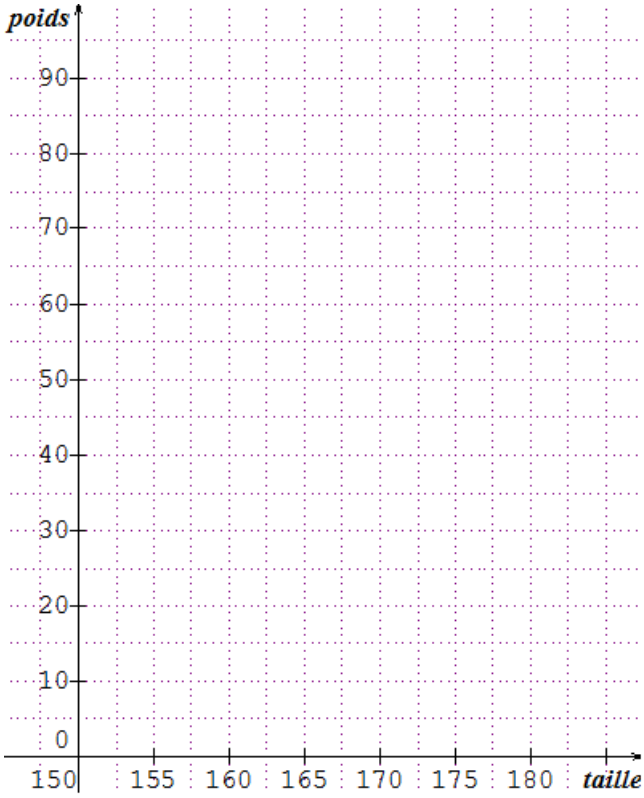
Activité d'introduction

Ce chapitre de statistique est présenté avec l'objectif de promouvoir la réflexion des étudiants et leur faire remarquer que ces notions statistiques sont somme toutes assez *'naturelles'*. Dans ce sens il élude à dessein le côté technique afin d'aller à l'essentiel. Il est donc vivement conseillé aux étudiants de s'investir pleinement dans les exercices: le chemin est tout aussi important que la destination.

Exercice 1

Voici les données donnant le poids et la taille des élèves d'une classe d'école de culture générale.

élève No	taille [cm]	poids [kg]
1	158	51
2	178	65
3	170	67
4	173	59
5	164	53
6	162	51
7	169	62.5
8	177	70
9	155	45



a) Que constatez-vous?

b) Représenter graphiquement ces données sur le repère ci-dessus. On appelle un tel graphique un graphique en nuage de points

c) Que constatez-vous à la vision de ce graphique?

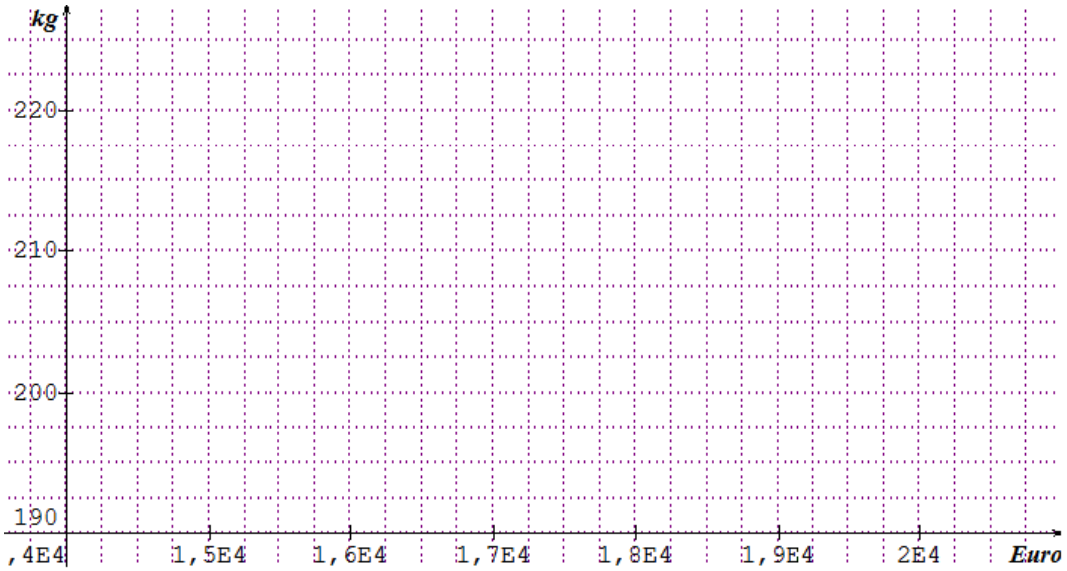
d) Comment expliquez-vous cela?

Exercice 2

Lors d'un test comparatif de motos sportives, on a recueilli les données suivantes :

Marque	**Modèle**	**Prix (€)**	**Puissance max (ch)**	**Couple maxi (m.kg)**	**Poids (kg)**
Suzuki	GSX-R 1000	14 299	185	12	208
Honda	CBR 1000 RR	15 690	175	11,7	199
KTM	1190 RC8 R	16 490	170	12,5	202
Aprilia	RSV4 Factory	19 990	180	11,7	206
MV Agusta	1000 F4	18 500	186	11,4	212
BMW	S 1000 RR	17 490	193	11,4	209
Yamaha	1000 R1	15 999	182	11,8	216
Kawasaki	ZX-10 R	14 599	188	11,5	208

a) Représenter le nuages de points sur le repère ci-dessous.



b) Que constatez-vous à la vision de ce graphique?

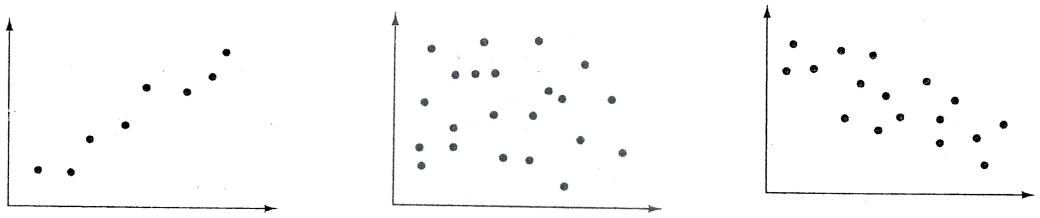
c) Comment expliquez-vous cela?

Corrélation

Exercice 3

Caractériser les graphiques en nuages de points ci-dessous.

1) 2) 3)



Exercice 4

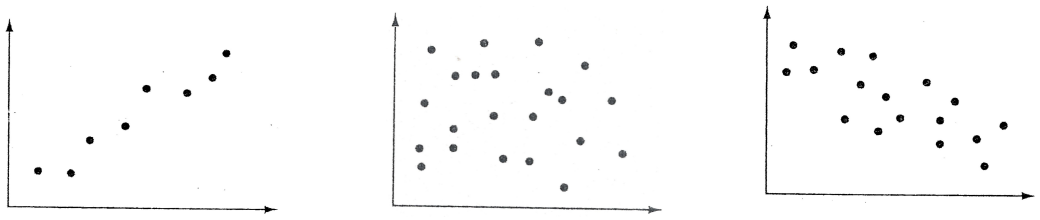
Pour les nuages de points ci-dessous:

a) Entourer les points d'un trait pour dessiner un nuage.

b) Si le nuage a une direction, dessiner une droite donnant la direction du nuage. On appelle cette droite la **droite de régression**.

c) Dire si les points sont plutôt proches ou pas de la droite.

1) 2) 3)



Exercice 5

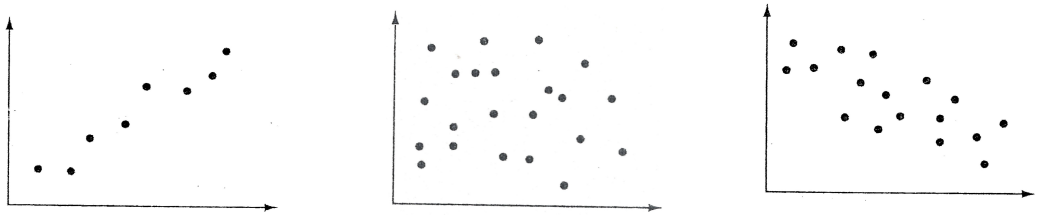
Pour les nuages de points ci-dessous:

a) Mettre les points en boite : c'est-à-dire dessiner le plus petit rectangle (oblique ou non) contenant tous les points du nuage.

b) Si la boîte a un côté plus large que l'autre et une direction oblique : dessiner une droite partageant le rectangle dans le sens de la longueur et donnant la direction du nuage. On appelle cette droite la **droite de régression**.

c) Dire si les points sont plutôt proches ou plutôt éloignés de la droite.

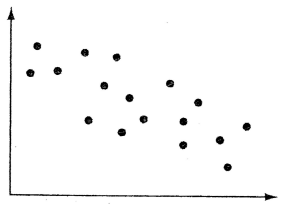
1) 2) 3)



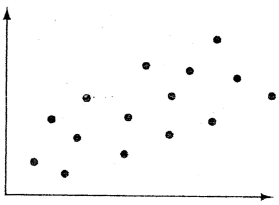
Exercice 6

Lequel des deux graphiques ci-dessous représente vraisemblablement les données de tailles et de poids des élèves d'une classe d'école?

1)



2)

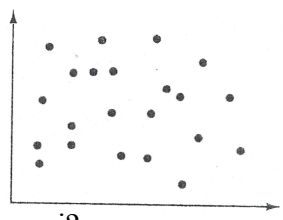


Pourquoi?

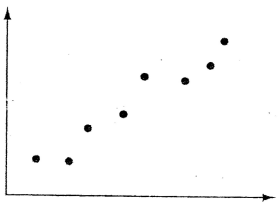
Exercice 7

Lequel des deux graphiques ci-dessous représente vraisemblablement les données d'âge et de nombre de points obtenus à un examen de mathématiques de l'ECG Adulte?

1)



2)



Pourquoi?

Théorie: corrélation et représentation graphique

Il y a **corrélation** entre deux grandeurs variables X et Y quand les deux grandeurs varient ensemble. On dira que la corrélation est:

- **positive** si la grandeur Y augmente lorsque la grandeur X augmente.
- **négative** si la grandeur Y diminue lorsque la grandeur X augmente.

Lorsque on représente les grandeurs X et Y sur un graphique en nuage de points, il y a corrélation lorsque le nuage de points a une direction ou la boîte a une direction oblique (ni horizontale ni verticale). Cette direction peut être représentée par une droite appelée **droite de régression**.

On dit que la corrélation est:

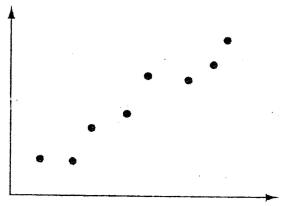
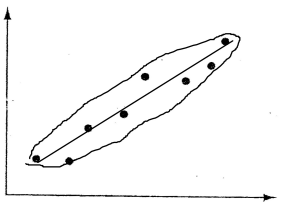
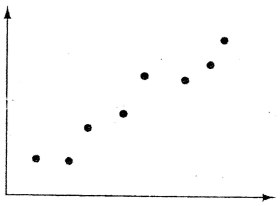
- **forte** si les points du nuage sont très proches de la droite de régression (boîte très étroite).
- **faible** si les points du nuage sont assez proches de la droite de régression (boîte assez étroite).

Dans ces deux cas la corrélation est dite **acceptable**.

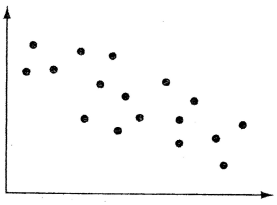
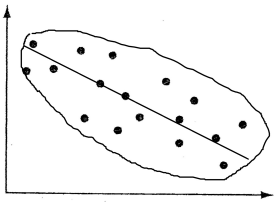
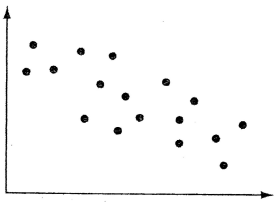
Il n'y a **pas de corrélation** (on dit aussi que la corrélation n'est pas acceptable) si le nuage de points n'a pas de direction claire.

Exemples:

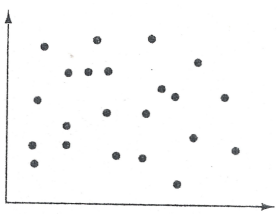
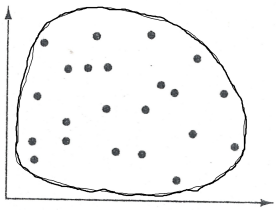
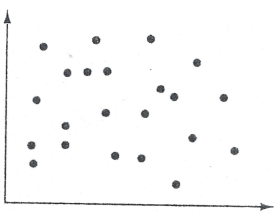
1) corrélation positive forte : nuage ou boîte proche de la droite de régression

2) corrélation négative faible : nuage ou boîte éloigné de la droite de régression

3) pas de corrélation : nuage sans 'direction' ou boîte 'carrée' ou horizontale
pas de droite de régression **pas** de droite de régression

Exercice 8

Pour les nuages de points ci-dessous déterminer :

a) S'il y a corrélation ou non et justifier votre réponse.

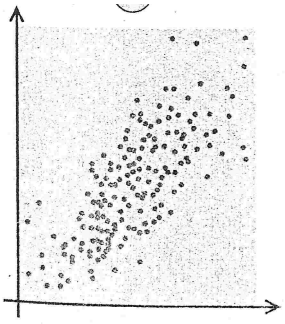
b) S'il y a corrélation :

i) Déterminer sa nature (si elle est positive ou négative).

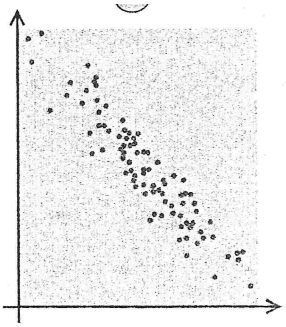
ii) Déterminer sa force (si elle est forte ou faible).

iii) Dessiner la droite de régression (et le nuage ou la boîte).

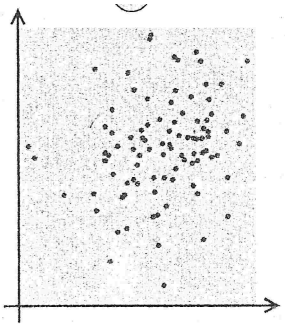
1



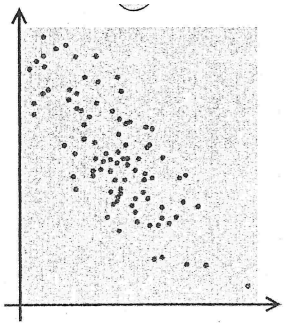
2



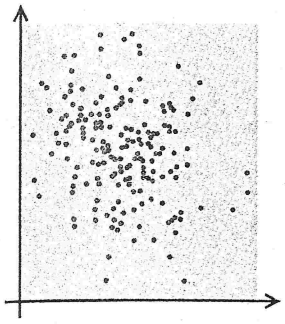
3



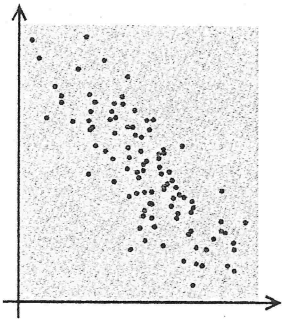
4



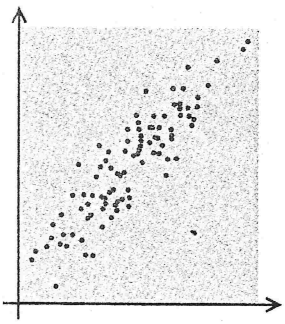
5



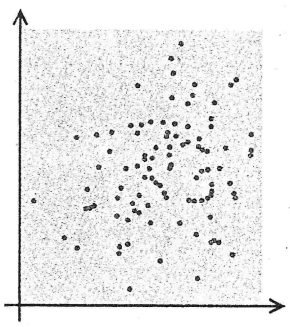
6



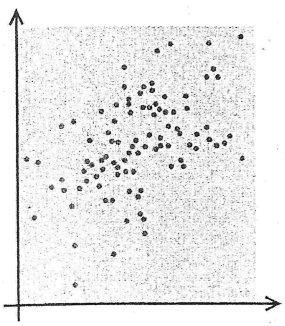
7



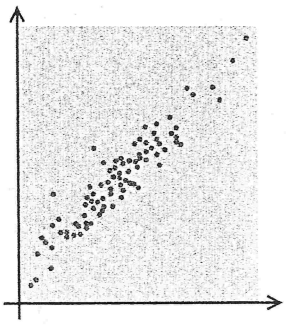
8



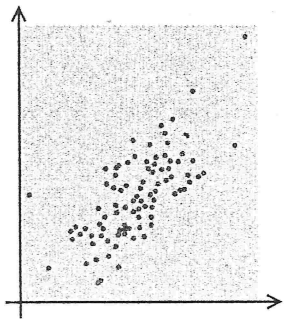
9



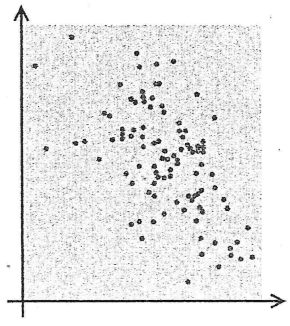
10



11



12



Causalité ?

Exercice 9

Est ce que le poids et la taille des gens sont corrélés?

Pourquoi?

Exercice 10

Il y a une forte corrélation positive entre le nombre d'accidents de chasse et le nombre de cartables vendus.

a) Expliquer ce que cela signifie.

b) Est-ce que l'un est la cause de l'autre ou y a-t-il une autre explication?

Exercice 11

Il y a une corrélation négative entre le nombre de cambriolages par commune et le nombre de couleurs du drapeau de la commune.

a) Expliquer ce que cela signifie?

b) Est-ce que l'un est la cause de l'autre ou y a-t-il une autre explication?

Exercice 12

Un chercheur en économie n'a pas trouvé de corrélation entre fortune (mesurée en dollars) et satisfaction amoureuse (mesurée sur une échelle de 1 à 10).
Peut-il y avoir un lien de cause à effet entre ces deux grandeurs?

Théorie: corrélation et causalité

Lorsque il y a corrélation entre deux grandeurs X et Y , cela signifie qu'il y a une relation de variabilité commune (les deux grandeurs augmentent simultanément ou l'une augmente lorsque l'autre diminue). Cependant corrélation ne signifie **pas** qu'il y ait un lien de cause à effet. En effet les cas suivants peuvent se produire:

- X cause Y
- X est causé par Y
- une autre grandeur Z cause simultanément X et Y ;
  c'est donc Z qui explique la corrélation de X et Y
- X ne cause pas et n'est pas causé par Y

Comme le quatrième cas est tout aussi possible que les trois autres, trouver qu'il y a corrélation entre deux grandeurs **ne permet pas de prouver un lien de cause à effet** mais permet seulement dire que les deux grandeurs varient ensemble.

Exercice 13

a) Un chercheur a trouvé qu'il n'y a pas de corrélation entre la pression artérielle et le degré de myopie d'une personne.
Peut-on dire s'il y a une relation ou pas entre ces deux grandeurs ?

Peut-on dire s'il y a une relation de cause à effet ou pas entre ces deux grandeurs ?

b) Deux grandeurs ne sont pas corrélées.
Peut-on affirmer qu'il n'y a pas de relation entre ces deux grandeurs ?

Peut-on affirmer qu'il n'y a pas de relation de cause à effet entre ces deux grandeurs ?

Exercice 14

a) Un chercheur a trouvé une forte corrélation positive entre le nombre de cigarettes fumées par jour et l'occurrence de cancer des poumons
Quelle grandeur cause l'autre ?

Pourquoi ? Comment ?

b) Un chercheur a trouvé une corrélation négative entre le nombre de cigarettes fumées par jour et l'occurrence de cancer du colon.
Quelle grandeur cause l'autre ?

Pourquoi ? Comment ?

c) Deux grandeurs sont fortement corrélées.
Peut-on affirmer qu'il y a une relation de cause à effet entre ces deux grandeurs ?

Exercice 15

Un chercheur a trouvé une corrélation moyenne entre le nombre de cigognes nichant sur les toits d'un village et le nombre de naissances dans le village.

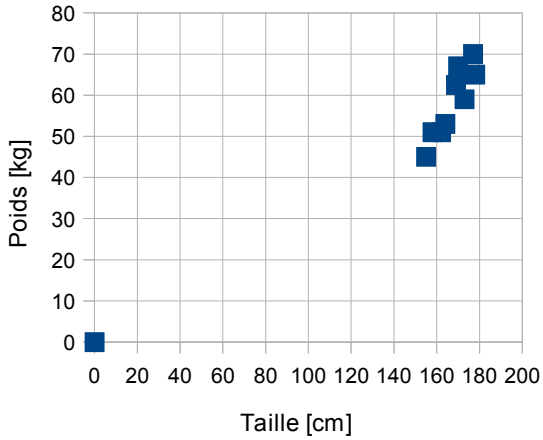
a) Expliquer ce que cela signifie.

b) Y a-t-il un lien de cause à effet?
Et si oui lequel?

Inférence statistique

Exercice 16

On a mesuré la taille et le poids des élèves d'une classe d'école. Les données sont représentées sur le graphique en nuage de points ci-dessous.



a) Y a-t-il corrélation entre ces deux grandeurs?

Justifiez votre réponse en fonction du graphique!

b) Dessiner la droite de régression.

c) Un élève de la classe, Alban, était absent lors des mesures, mais on sait qu'il mesure 168 cm.

i) Inférer (prédire, estimer en fonction des données déjà connues) le poids d'Alban.

ii) Dessiner sur le graphique ci-dessous le point A représentant Alban.

d) Une autre élève de la classe, Béatrix, était aussi absente lors des mesures, mais on sait qu'elle pèse 65kg.

i) Inférer la taille de Béatrix.

ii) Dessiner sur le graphique ci-dessous le point B représentant Béatrix.

Exercice 17

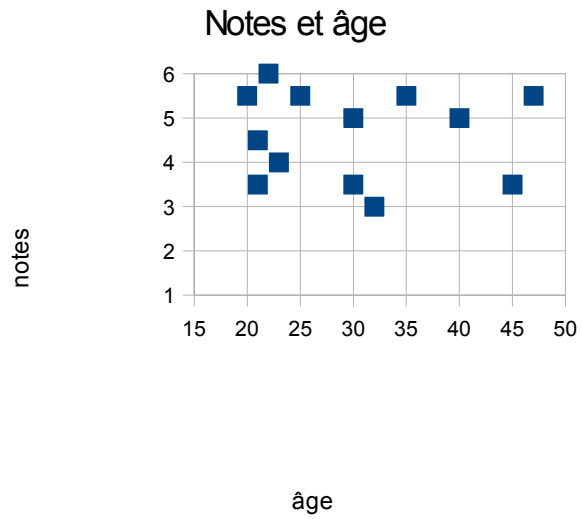
Le graphique ci-dessous représente l'âge et la note d'examen de mathématiques obtenue par les élèves d'une classe au collège du soir.

a) Y a-t-il corrélation entre ces deux grandeurs?

Justifiez votre réponse en fonction du graphique!

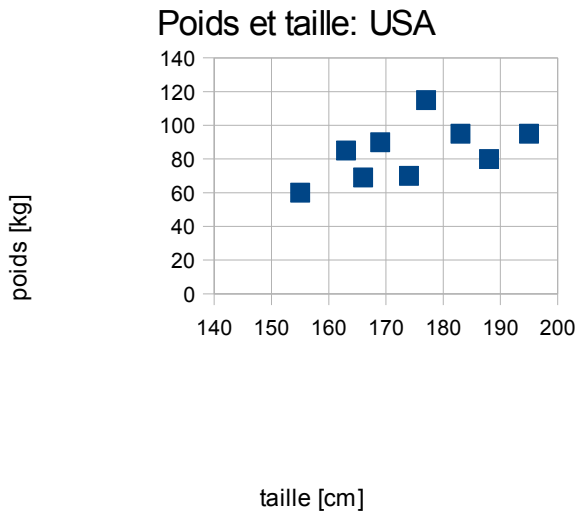
b) Est-il possible de dessiner la droite de régression?

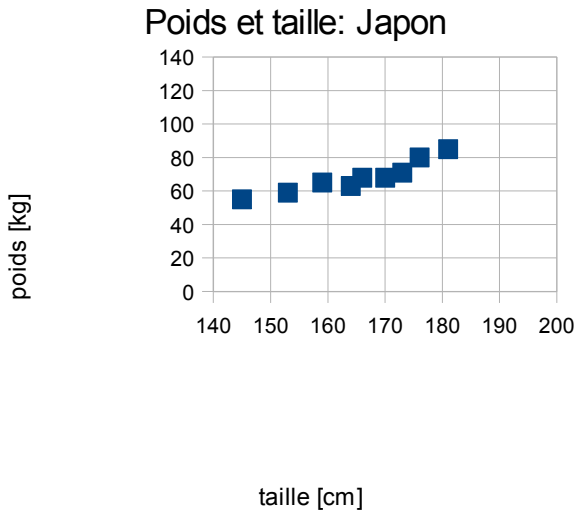
c) Un élève de la classe, Casimir, était absent lors de l'examen. Est-il possible d'inférer la note qu'il aurait eu sachant qu'il a 37 ans.



Exercice 18

Les deux graphiques ci-dessous représentent la taille et le poids de deux échantillons de touristes: un provenant des Etats-Unis, l'autre du Japon.





a) A partir de leur représentation graphique, comparer ces deux échantillons.

b) Y-a-il corrélation? Est-elle forte ou faible?

USA: Japon:

c) Inférer le poids *d'un touriste de ces mêmes groupes* sachant qu'il mesure 175 cm.

USA: Japon:

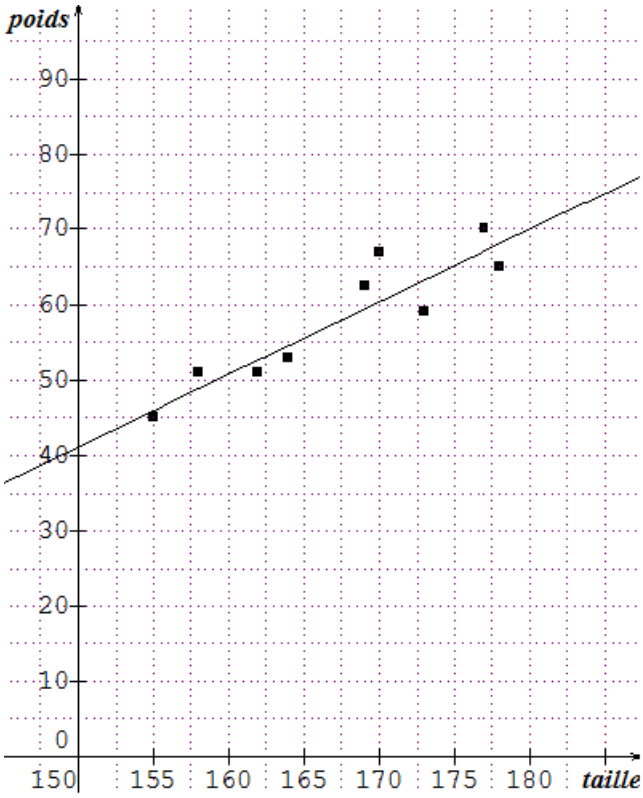
d) Qu'en est-il de la qualité de votre prédiction?

USA: Japon:

e) Peut on inférer le poids d'un japonais de 175 cm qui *n'est pas membre de ce groupe de touristes*?

Exercice 19

Voici le graphique représentant la taille et le poids d'un échantillon de personnes.
Il y a corrélation entre les deux variables et la droite de régression à été dessinée.



a) Utiliser la droite de régression pour inférer le poids d'une personne de

i) 170 cm.

ii) 100 cm

iii) 300 cm

b) Que remarquez vous?

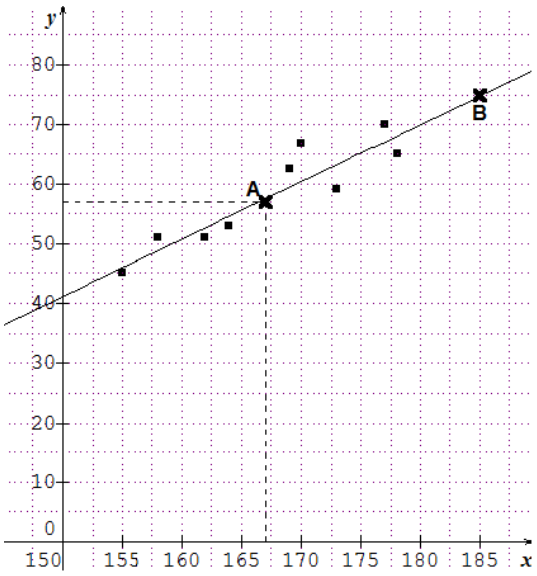
c) Que peut-on conclure sur la validité de nos prédictions?

Théorie: inférence statistique

Inférer signifie prédire en fonction des données connues en utilisant un outil statistique.

Si deux grandeurs variables d'une population sont corrélées,
Alors on peut inférer (prédire) une des grandeur d'un individu de la population en connaissant l'autre grandeur le concernant.

Pour cela on utilise la droite de régression: la grandeur connue (x ou y) nous permet de dessiner un point (A) sur la droite de régression qui nous permet ensuite de lire la grandeur inconnue (y ou x).



Remarques: Il faut toutefois faire attention aux considérations suivantes:

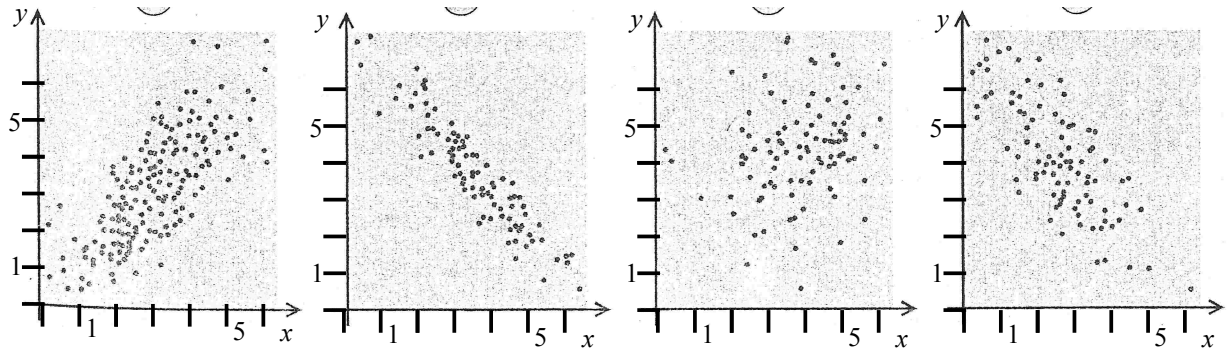
- La droite de régression nous permet de faire la meilleure prédiction possible en fonction des données connues. Cependant ce n'est qu'une prédiction: la réalité peut être différente!
- La qualité de la prédiction va dépendre de la force de la corrélation:
  - plus la corrélation est forte et plus la prédiction devrait être fiable,
  - plus la corrélation est faible et moins la prédiction est fiable.

  Dans le cas d'une corrélation très faible on ne se permettra même pas de faire de prédiction car cette dernière ne peut pas être assez fiable pour avoir du sens.
- Les prédictions doivent être limitées à des individus qui se trouveraient dans le nuage de points. (C'est-à-dire que les valeurs connues doivent être comprises entres le minimum et le maximum des valeurs de l'échantillon.) Ainsi on peut faire une prédiction pour A mais pas pour B!
- Evidemment la qualité de l'échantillon aura une influence sur la qualité de la prédiction: plus l'échantillon est représentatif de la population et plus la prédiction sera fiable. Les deux plus importants facteurs influençant la qualité de l'échantillon sont: la taille échantillon (le plus grand sera le mieux) et la méthode de sélection de l'échantillon (aléatoire, contrôle, etc.).

Exercice 20

Pour les graphiques en nuages de points ci-dessous inférer si possible la valeur de y sachant que x vaut 5. Justifier votre réponse.

1 2 3 4



Exercice 21

a) Reprendre les prédictions que vous avez faites à l'exercice précédent et donner maintenant vos réponses sous forme de fourchette pour tenir compte de la variabilité et donc de la fiabilité de votre prédiction.

b) Quel lien y a-t-il entre fiabilité de la prédiction et la force de la correlation ?

Exercice 22

Voici l'âge et le pourcentage d'acuité auditive de cinq personnes âgées.

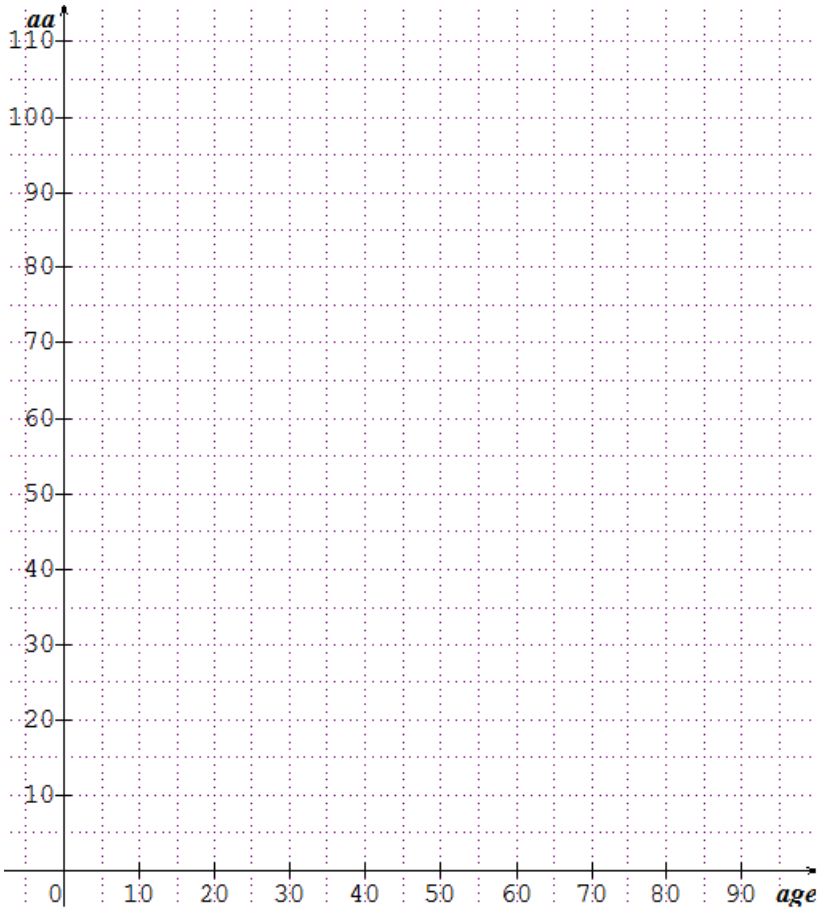
Age (ans)	Acuité (%)
65	97
66	80
70	92
75	78
80	70

a) Dessiner le graphique en nuage de points représentant ces données.
b) Interpréter la situation à l'aide du graphique.

c) Y a-t-il corrélation?
Si oui donner sa nature.

d) Peut-on inférer l'âge d'une personne ayant 75% d'acuité auditive ?
Si oui, justifiez et faites le !
Si non, justifiez !

e) Peut-on prédire l'acuité auditive d'une personne de 50 ans ?
Si oui, justifiez et faites le !
Si non, justifiez !



Objectifs du cours et auto-évaluation

A la fin de ce cours, et dans le but d'être prêts pour l'examen, vous devez:

savoir

- la définition d'une corrélation (positive ou négative)
- qu'une corrélation n'implique pas de lien de cause à effets

être capable de

- dessiner un graphique en nuage de points
- déterminer à partir d'un graphique en nuage de points s'il y a corrélation ou non et si oui : si elle est positive ou négative, faible ou forte ;
- dessiner approximativement sur un graphique en nuage de points:
  - la boite ou nuage encerclant les points au mieux
  - la droite de régression
  - un point supplémentaire en ne connaissant qu'une de ses données (x ou y) et en déduire l'autre donnée (y respectivement x)

Table des matières