Remarque : Si les listes contiennent déjà des données, il faut les effacer.

Appuyer sur Stat et sélectionner «4 : ClrList». Ensuite, sélectionner les colonnes à effacer.
Par exemple, pour effacer les colonnes L1 et L2, il faudra faire : 2nd 1 , 2nd 2 ENTER

```
ClrList  L1, L2
              Done
```

B) Calcul du coefficient de corrélation linéaire

Entrer les listes. Pour calculer le coefficient de corrélation linéaire, appuyer sur STAT et déplacer le curseur pour choisir le menu CALC.

TI 83

```
EDIT CALC TESTS
1 : 1-Var Stats
2 : 2-Var Stats
3 : Med-Med
4 : LinReg (ax + b)
5 : QuadReg
6 : CubicReg
7  QuartReg
```

Pour la TI 83, sélectionner l'option «4 : LinReg (ax + b)».

TI 82

```
EDIT CALC
1 : 1-Var Stats
2 : 2-Var Stats
3 : SetUp...
4 : Med-Med
5 : LinReg (ax + b)
6 : QuadReg
7  CubicReg
```

Pour la TI 82, sélectionner l'option «5 : LinReg (ax + b)».

Les deux calculatrices affichent :

```
LinReg (ax + b)
```

Pour choisir les variables (L1 et L2) dont on veut calculer le coefficient de corrélation linéaire, appuyer sur [2nd] [1] [,] [2nd] [2] [ENTER]

Sur la TI 82, les paramètres de la droite de régression ainsi que le coefficient de corrélation linéaire apparaissent à l'écran.

```
LINREG (a + bx)
y = ax + b
a = 9,238070806
b = 10,37660339
r = .9940016268
```

Sur la TI 83, seulement les paramètres de la droite de régression apparaissent. Pour afficher le coefficient de corrélation « r », appuyer sur [VARS] et sélectionner «5 : statistics...».

Ensuite, déplacer le curseur pour choisir le menu EQ. Sélectionner «7 r» puis appuyer sur [ENTER].

```
VARS Y-VARS
1 : Windiw...
2 : Zoom...
3 : GDB...
4 : Picture...
5: Statistics...
6: Table...
7: String...
```

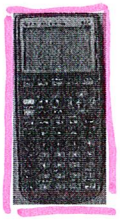
```
XY Σ EQ TEST PTS
1 : RegEQ
2 : a
3 : b
4 : c
5 : d
6 : e
7 r
```

```
a = 9,238070806
b = 10,37660339


r

          .9940016268
```

Exemple :



Le tableau ci-contre fournit de l'information sur le début de la saison des Canadiens de Montréal en 2008-2009.

Joueur	Nombre de matchs disputés	Points produits
Francis Bouillon	36	6
Alex Kovalev	41	32
Andrei Markov	41	31
Alex Tanguay	34	26
Andrei Kostitsyn	37	25
Saku Koivu	28	22
Tomas Plekanec	41	20
Sergei Kostitsyn	38	18
Guillaume Latendresse	34	16
Roman Hamrlik	40	15
Maxim Lapierre	38	13
Patrice Brisebois	37	12
Steve Bégin	31	9
Georges Laraque	17	2
Josh Gorges	41	8

Peut-on affirmer qu'il existe une corrélation linéaire entre le nombre de matchs disputés et le nombre de points produits ?

1) Entrer les listes

2) Calculs a et b $a \approx 0,6413$ $b \approx -5,8315$

3) Calcul "r"

$r \approx 0,4585$

* Regarder le nuage de points.

Réponse : Pas vraiment, le coefficient de corrélation linéaire est trop faible.

4. La droite de régression

Lorsque la corrélation linéaire est suffisamment forte, on observe un lien de dépendance entre les variables statistiques étudiées. Dans ce cas, il devient possible de **modéliser** la situation par une droite.

Définition : La **droite de régression** est la droite qui s'ajuste le mieux à un nuage de points présentant une corrélation linéaire.

* (Estimer)
Il existe différentes méthodes pour déterminer l'équation d'une droite de régression.

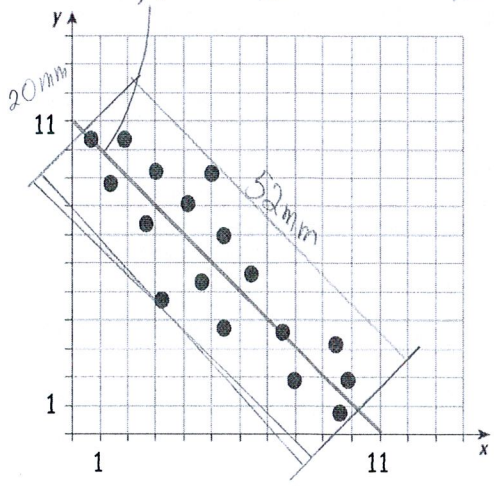
A) À l'aide d'une droite tracée à main levée.
B) À l'aide de la calcultrice.
C) À l'aide de la méthode de la droite de Mayer.
D) À l'aide de la méthode de la droite médiane-médiane.
E) Autres, non vues cette année

A) Une droite tracée à main levée

On aurait pu tracer la droite nous-même.

Exemple :

Soit le nuage de points ci-contre :



a) Estimer graphiquement le coefficient de corrélation linéaire.

$r \approx \pm \left(1 - \frac{p}{q}\right)$

$r \approx -\left(1 - \frac{20}{52}\right)$

$r \approx -0,62$

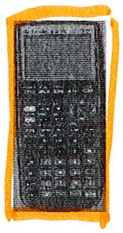
Le coefficient est :

b) Déterminer l'équation de la droite de régression. (0, 11) (11, 0)

$a = \frac{y_2 - y_1}{x_2 - x_1} = \frac{0-11}{11-0} = \frac{-11}{11} = -1$

L'équation de la droite de régression est : $y = -x + 11$

B) La droite de régression et la calculatrice



Revenons à l'exemple de la page 216 du manuel.

Le coefficient de corrélation linéaire nous indique qu'il existe une corrélation linéaire positive et très forte entre l'âge d'un jeune baobab et la mesure de sa circonférence.

Dans ce cas, il est possible de **modéliser** la situation par une **droite de régression** et d'utiliser son équation afin de faire des prédictions concernant l'âge du baobab et sa circonférence.

Ces paramètres sont :

a = 9,238070806
b = 10,37660339

Donc l'équation de la droite de régression est :

y = 9,2381x + 10,3766 où x représente l'âge du baobab en années
y représente la circonférence en cm

→ Obtenir le graphique et la table de valeurs :

Appuyer sur [y=] pour avoir accès à une équation. Appuyer sur [VARS] et sélectionner «5 : statistics...» pour la TI 83 et la TI 82 (ou sélectionner «2 STATISTICS...» pour la TI 80.) Déplacer le curseur pour choisir le menu EQ.

TI 83	TI 82
XY Σ EQ TEST PTS	X/Y Σ EQ BOX PTS
1 : RegEQ	1 : a
2 : a	2 : b
3 : b	3 : c
4 : c	4 : d
5 : d	5 : e
6 : e	6 : r
7 r	7 : RegEQ
Sélectionner «1 :RegEQ».	Sélectionner «7 : RegEQ».

La calculatrice affiche alors l'équation de la droite de régression dans la fenêtre des équations.

Appuyer sur [GRAPH], pour tracer la droite de régression.

Pour avoir accès à la table de valeurs, appuyer sur [2nd] [GRAPH].

C) Méthode de la droite de Mayer

La droite de Mayer est la droite passant par **2 points moyens P_1 et P_2** représentatifs de la distribution.

Procédure :

1) ORDONNER les données en ordre croissant selon la 1ere variable (x) puis PARTAGER la distribution en 2 groupes égaux.

   - Pour **2 valeurs égales de x**, ordonner les valeurs de y en ordre croissant (corrélation positive) ou ordre décroissant (corrélation négative).
   - Si le **nombre de données est impair**, la donnée du centre est placée dans chacun des deux groupes.

2) Trouver les coordonnées de $P_1(x_1, y_1)$ de la 1ere moitié de la distribution.

   x_1 = moyenne des abscisses y_1 = moyenne des ordonnées

3) Trouver les coordonnées de $P_2(x_2, y_2)$ en procédant de la même manière avec la deuxième partie.

4) Trouver l'équation $y = ax + b$ de la droite PASSANT par P_1 et P_2.

Impair → donnée du milieu dans le 2 sous-groupes.

Exemple : Déterminer l'équation de la **droite de Mayer** à partir des données de la distribution suivante.

x	~~14~~	~~6~~	~~22~~	~~27~~	~~14~~	~~3~~	~~9~~	19	~~18~~	~~15~~	~~16~~ 12
y	6	5	21	17	10	2	7	14	9	12	

1) ordre

x	3	6	9	14	14	15	16	18	19	22	27
y	2	5	7	6	10	12	12	9	14	21	17

↑ 2 groupes

2) $P_1(x_1, y_1)$

$x_1 = \frac{61}{6}$ $\quad y_1 = \frac{42}{6}$

$x_1 \approx 10,17$ $\quad y_1 = \frac{42}{6}$

$P_1(10,17 ; 7)$

3) $P_2(x_2, y_2)$

$x_2 = \frac{117}{6}$ $\quad y_2 = \frac{85}{6}$

$x_2 = 19,5$ $\quad y_2 \approx 14,17$

$P_2(19,5 ; 14,17)$

3) Droite de régression

$a = \frac{y_2 - y_1}{x_2 - x_1}$

$a = \frac{14,17 - 7}{19,5 - 10,17}$

$a \approx \frac{7,17}{9,33}$

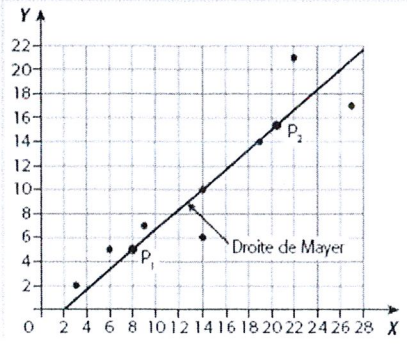
$a \approx 0,77$

$y = 0,77x + b$

$7 = 0,77 \cdot 10,17 + b$

$7 = 7,83 + b$

$-0,83 = b$



L'équation de la droite de Mayer est $y = 0,77x - 0,83$

D) Méthode de la droite médiane-médiane

La droite médiane-médiane est la droite définie à partir de **3 points médians M_1, M_2 et M_3** représentatifs de la distribution.

Procédure :

1) ORDONNER les données en ordre croissant selon x.

   ⇨ Pour 2 valeurs égales de x, ordonner les valeurs de y en ordre croissant (corrélation positive) ou ordre décroissant (corrélation négative).

2) PARTAGER la distribution en 3 groupes.
   À l'intérieur de chaque groupe, replacer les valeurs de y en
   - ordre croissant (r+)
   - ordre décroissant (r−)

   ⇨ Le 1er et le 3e groupe doivent être équipotents. (même nb données)

   ⇨ Les 3 groupes doivent être le plus possible équipotents

3) Pour le premier groupe, déterminer $M_1(x_1, y_1)$
   x_1 = médiane des abscisses y_1 = médiane des ordonnées

4) Pour le deuxième groupe, déterminer $M_2(x_2, y_2)$.

5) Pour le troisième groupe, déterminer $M_3(x_3, y_3)$.

6) Déterminer les coordonnées du point P à partir de M_1, M_2, M_3.

$$x = \frac{x_1+x_2+x_3}{3} \quad \text{et} \quad y = \frac{y_1+y_2+y_3}{3}$$

7) Trouver l'équation $y = ax + b$ sachant que :
   - Elle passe par le point P
   - Elle est parallèle à la droite passant par M_1 et M_2.

Exemple : Déterminer l'équation de la **droite médiane-médiane** à partir des données de la distribution suivante.

x	14	6	22	27	14	3	9	19	18	15
y	6	5	21	17	10	2	7	14	9	12

1) ORDRE (3 groupes + replacer y)

x	3	6	9	14	14	15	18	19	22	27
y	2	5	7	6	9	10	12	14	21	17

M_1 M_2 M_3

2) $M_1(6, 5)$

3) M_2

$x_2 = \frac{14+15}{2}$ $y_2 = \frac{9+10}{2}$

$x_2 = 14,5$ $y_2 = 9,5$

$M_2(14,5 ; 9,5)$

4) $M_3(22, 17)$

5) Point P

$x = \frac{6+14,5+22}{3}$

$x = 14,17$

$y = \frac{5+9,5+17}{3}$

$y = 10,5$

$P(14,17 ; 10,5)$

6) Pente $\overline{M_1M_3}$

$m = \frac{y_2 - y_1}{x_2 - x_1}$

$m = \frac{17-5}{22-6}$

$m = \frac{12}{16} = 0,75$

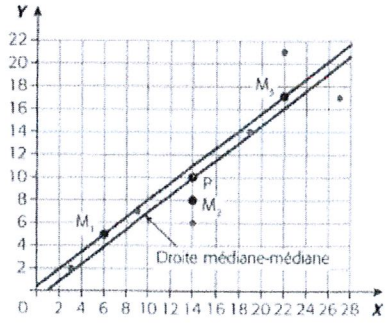
7) Equation Méd-Méd $(m = 0,75)$ $(14,17 ; 10,5)$

$y = 0,75x + b$

$10,5 = 0,75 \cdot 14,17 + b$

$-10,63 \quad -10,63$

$-0,13 = b$



L'équation de la droite médiane-médiane est $y = 0,75x - 0,13$

Remarques :

◊ Un coefficient de corrélation égal à 1 ou -1 indique une corrélation parfaite de l'échantillon (aucune différence entre les valeurs « y » estimées et réelles). À l'inverse, un coefficient de corrélation égal à 0 indique que **l'équation de régression** ne peut servir à prévoir une valeur « y ».

◊ Pour déterminer l'équation de la droite de régression, il est plus simple de déterminer celle de la droite de Mayer quand il y a peu de données et celle de la droite médiane-médiane quand il y en a beaucoup.

◊ Il est préférable d'avoir recours à l'équation de la droite médiane-médiane si la distribution présente des points aberrants.

E) Interpolation et extrapolation

Interpolation : Estimation d'une valeur située à l'**intérieur** du nuage de points.

Extrapolation : Estimation d'une valeur située à l'**extérieur** du nuage de points.

Remarques :

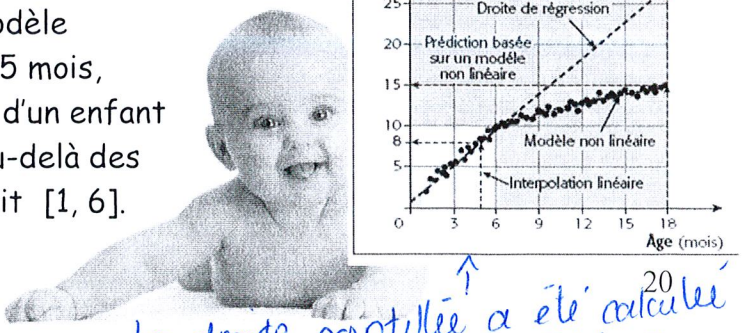
◊ Une prédiction par **interpolation** est généralement **plus fiable** qu'une prédiction par extrapolation.

◊ Avec une **extrapolation**, rien ne garantit que le modèle linéaire puisse être étendu à l'extérieur des limites de l'intervalle des données pour lesquelles il a été établi.

◊ Plus on s'éloigne de cet intervalle, plus le risque d'obtenir une prédiction aberrante est grand.

Exemple :

Dans le diagramme ci-contre, on peut voir que le modèle linéaire permet de prédire la masse d'un enfant de 5 mois, mais qu'il n'est pas approprié pour prédire la masse d'un enfant de 18 mois. En effet, le modèle ne s'applique pas au-delà des limites de l'intervalle pour lequel il a été établi, soit [1, 6].

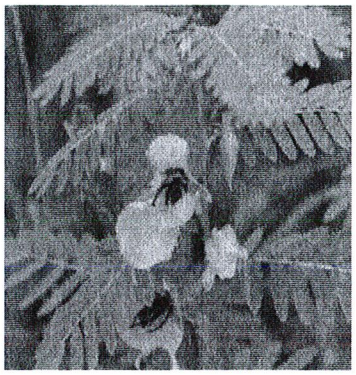


La droite pointillée a été calculée à partir des pts du carré blanc.

F) Prédiction à l'aide de la droite de régression

La modélisation de la corrélation linéaire par une droite de régression permet de prédire la valeur d'une des deux variables à partir de l'autre. En général, plus la corrélation linéaire est forte, plus la prédiction est fiable.

Exemple : Le tableau ci-dessous présente les mesures recueillis par un biologiste à propos de certaines caractéristiques d'une plante qu'il étudie.



Circonférence de la tige (cm)	Hauteur de la plante (cm)
0,3	8,3
0,4	10,2
0,45	10,3
0,65	13,5
0,8	13,8
0,85	15,1
1	16
1,1	18,1
1,15	18,6
1,2	21
1,3	20,6
1,4	22

Estimer la hauteur de la plante lorsque la circonférence sera de 2 cm.

Réponse : _____

5. Facteurs intervenant dans l'interprétation de la corrélation

Plusieurs facteurs peuvent intervenir dans l'interprétation de la corrélation entre deux variables. C'est pourquoi il faut être vigilant lorsque vient le temps de faire des prédictions et de tirer des conclusions.

Interprétation	Exemple
Le lien entre deux variables peut être un rapport de cause à effet, c'est-à-dire que l'une des variables agit directement sur l'autre.	La corrélation entre l'altitude et la température puisque la température varie directement en fonction de l'altitude.
La corrélation entre deux variables peut être importante sans que les deux variables soient directement liées entre elles. Elles peuvent dépendre toutes les deux d'une troisième variable qui, en variant, engendre des variations pour les deux premières.	En été, il peut sembler y avoir une forte corrélation entre le nombre de cornets de crème glacée vendus et le nombre de climatiseurs vendus dans une ville, alors qu'en fait, ces deux variables dépendent plutôt d'une troisième, qui est la température.
Considérer une corrélation comme étant linéaire alors qu'un autre modèle serait plus approprié.	La croissance de la population d'une métropole peut être étudiée selon une corrélation linéaire. Toutefois, l'utilisation d'un modèle exponentiel serait plus appropriée.
Il arrive parfois qu'il y ait une corrélation entre deux variables seulement sur un intervalle donné.	Sur l'intervalle $[5, 10]$ ans, la corrélation entre l'âge et la taille d'une personne est linéaire. Toutefois, avant et après cet intervalle, le modèle linéaire n'est pas le mieux adapté.
Une distribution à deux variables peut comporter des données aberrantes, en raison notamment d'erreurs de manipulation ou de mesure.	Le degré de précision de l'instrument utilisé lors de la collecte des données laisse à désirer.

Une forte corrélation indique l'existence d'un lien statistique. Cependant, elle n'explique pas la raison et la nature du lien. Par la suite, on essaie de caractériser qualitativement et quantitativement ce lien, et d'établir des prédictions tout en étant *conscient des limitations de ces prédictions*.