Chapitre VI : notions de statistiques appliquées aux mesures

"Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation."

Encyclopedia Universalis

I. Généralités sur la statistique

I.1 Variable statistique

Soit une grandeur physique *X* dont la valeur exacte est x_0.

n mesures conduisent à des valeurs x_1, x_2, … x_n

La grandeur *X* est appelée variable statistique. Les valeurs qu'elle peut prendre sont notées x_1, x_2, …

La variable statistique peut être de nature continue (longueur, masse, temps, concentration …) ou de nature discrète (nombre de défectueux dans un lot de fabrication, nombre d'individus possédant une caractéristique donnée …)

I. Généralités sur la statistique

I.2 But de la statistique appliquée aux mesures

Objectif : donner une estimation de la différence maximale entre la mesure x et la vraie valeur x_0.

On fixe au préalable un *risque d'erreur*, les résultats sont alors donnés en fonction de ce risque.

Pour un nombre infini de mesures (*sans présence d'erreur systématique*), on devrait en théorie obtenir la vraie valeur x_0.

La statistique : extrapole les résultats obtenus pour un nombre fini de mesures.

Echantillon : série limitée de résultats employés pour l'estimation

I. Généralités sur la statistique

I.3 Présentation des résultats. Histogrammes

Distribution statistique : mise en ordre des données observées.

→ on regroupe les résultats identiques ou appartenant à une même classe.

Valeur observée	x_1	x_2	…	…	…	x_n
Nombre d'observations	n_1	n_2	…	…	…	n_n

Les résultats sont présentés non pas sous forme de liste (peu pratique), mais sur un graphique appelé histogramme des effectifs.

I. Généralités sur la statistique

I.3 Présentation des résultats. Histogrammes

On regroupe les informations en classes : division de l'intervalle dans lequel le caractère *x* varie en intervalles plus petits $[x_1, x_2[$, $[x_2, x_3[$ …

- Les classes doivent être contiguës sans chevauchement (chaque observation doit pouvoir se mettre dans une classe sans ambiguïté)
- Le regroupement des valeurs revient à assimiler toutes les observations d'une même classe à un caractère unique : celui du point médian
- Perte d'informations d'autant plus grande que l'intervalle de la classe est étendu

I. Généralités sur la statistique

I.3 Présentation des résultats. Histogrammes

La distribution des effectifs est souvent représentée par un histogramme :

Chaque classe est représentée par un rectangle dont la base est proportionnelle à l'amplitude de la classe et la hauteur à l'effectif.

Histogramme = résumé des observations en un simple coup d'œil.

I. Généralités sur la statistique

I.3 Présentation des résultats. Histogrammes

Application : Vérification de la masse de paquets de tabac.

Prélèvement au hasard de 20 paquets soumis à la pesée. Les masses sont données en g.

41,0	40,4	40,6	40,7	41,2	41,2	41,5	41,6	42,3	42,9
39,2	39,8	40,5	40,2	42,1	41,6	41,3	40,9	40,7	40,1

Résultat de la pesée pour un échantillon de 20 paquets. Les masses sont données en g.

Tracer l'histogramme des effectifs pour des classes d'amplitude élémentaire 1 g

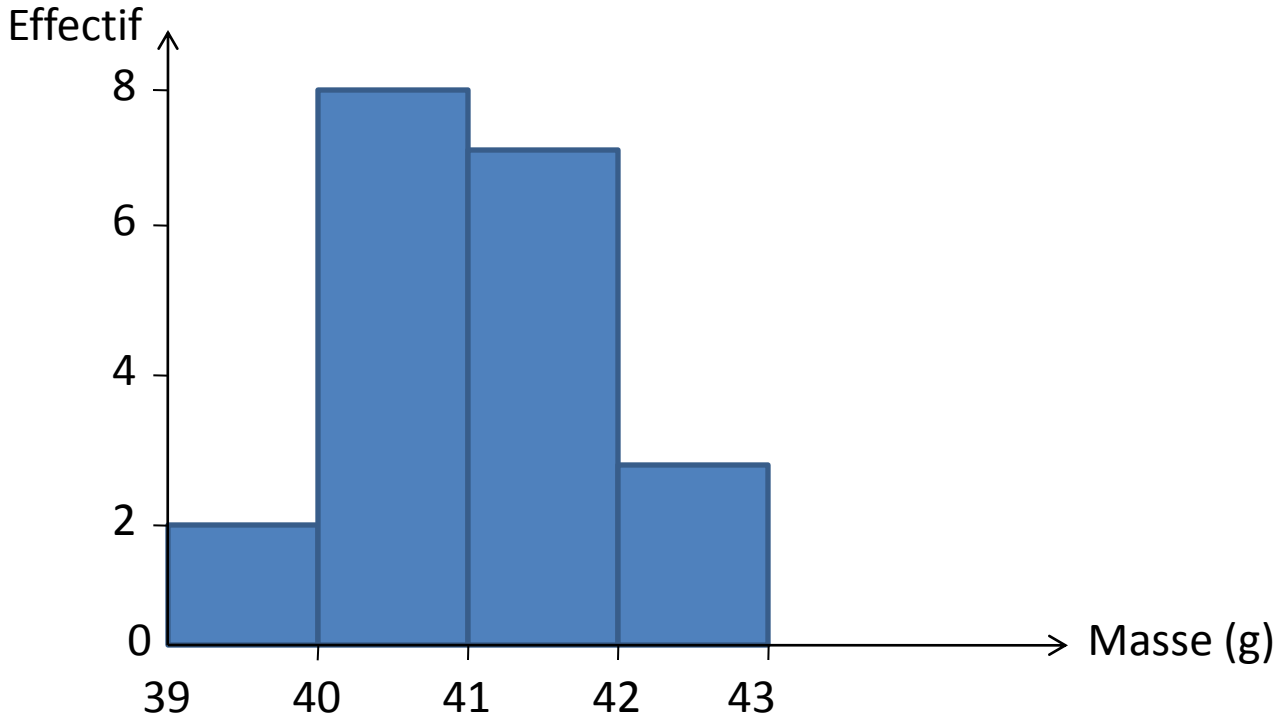
I. Généralités sur la statistique

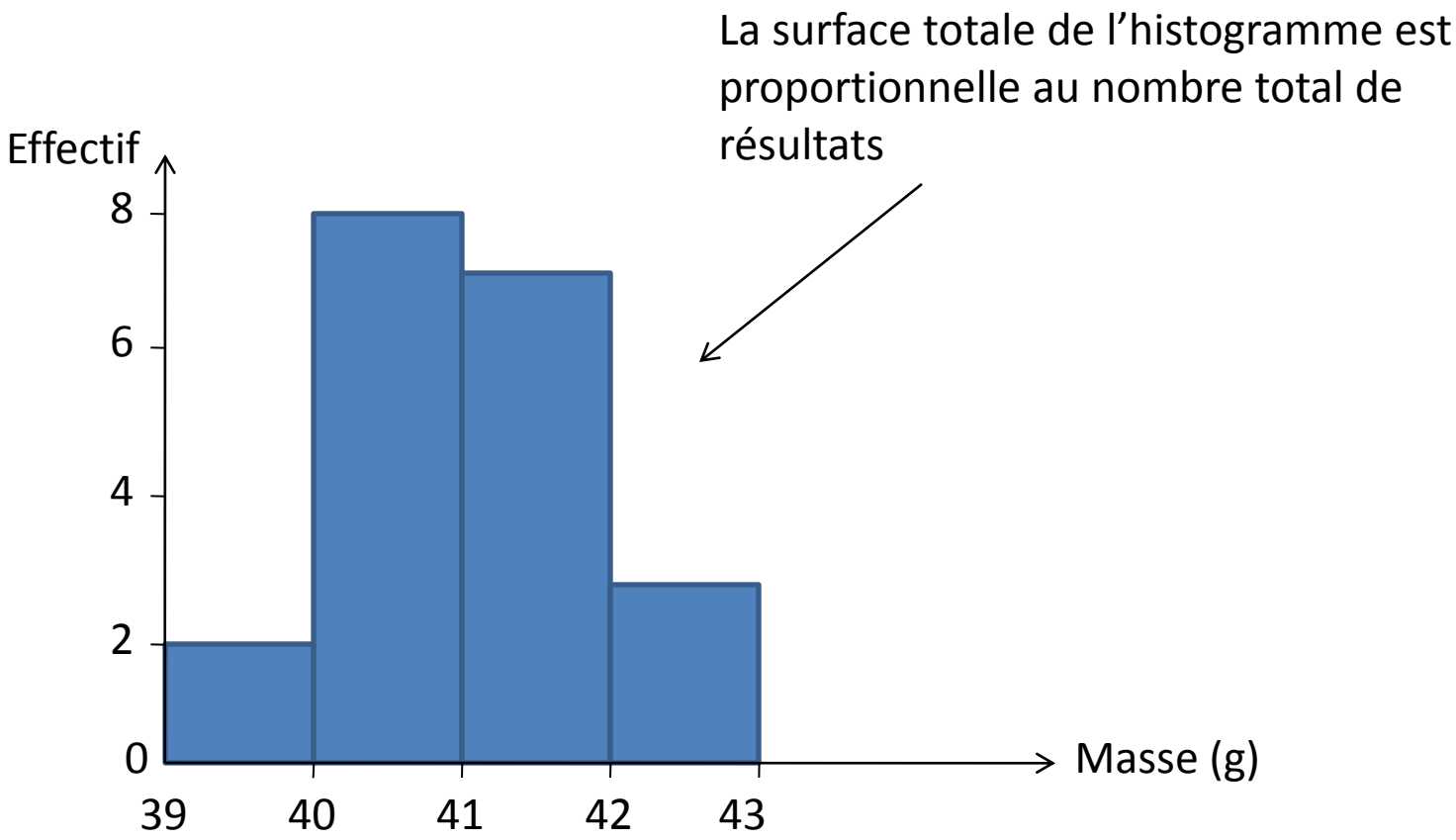
I.3 Présentation des résultats. Histogrammes

41,0	40,4	40,6	40,7	41,2	41,2	41,5	41,6	42,3	42,9
39,2	39,8	40,5	40,2	42,1	41,6	41,3	40,9	40,7	40,1

On range les échantillons dans des classes de 1 g entre 39 g et 43 g

[39 ; 40[2
[40 ; 41[8
[41 ; 42[7
[42 ; 43[3





Si les classes ne sont pas de même amplitude (par exemple classes de 0,5 g et classes de 1 g), alors il faut garder la proportionnalité entre surface et nombre de résultats)

$\Rightarrow$ on se ramène à la plus petite amplitude
$\Rightarrow$ on divise la hauteur du rectangle par le rapport de l'amplitude de la classe à l'amplitude élémentaire. Autrement dit, une classe deux fois plus grandes, aura une hauteur deux plus petite.

Application : prendre les classes suivantes et tracer l'histogramme correspondant :

[39 ; 40[

[40 ; 40,5[

[40,5; 41[

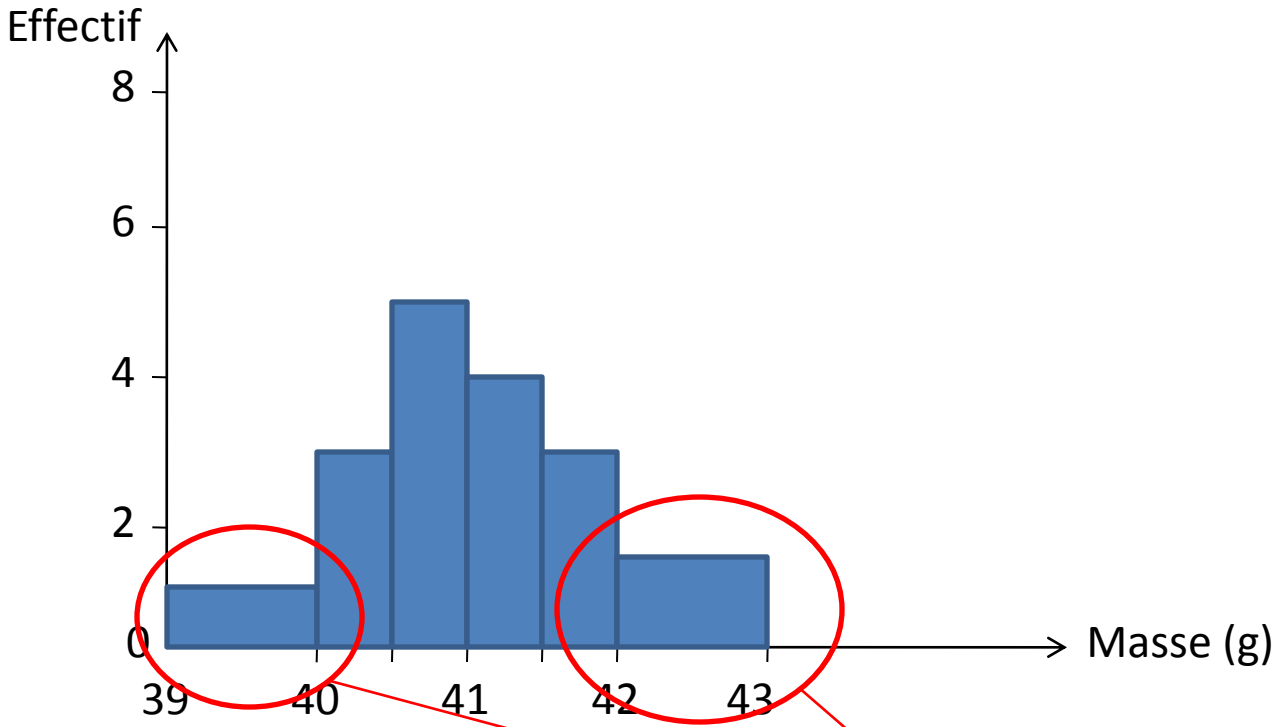
[41 ; 41,5[

[41,5 ; 42[

[42 ; 43[

41,0	40,4	40,6	40,7	41,2	41,2	41,5	41,6	42,3	42,9
39,2	39,8	40,5	40,2	42,1	41,6	41,3	40,9	40,7	40,1

Résultat de la pesée pour un échantillon de 20 paquets. Les masses sont données en g.



[39 ; 40[2
[40 ; 40,5[3
[40,5; 41[5
[41 ; 41,5[4
[41,5 ; 42[3
[42 ; 43[3

Classes 2 fois plus larges que la classe élémentaire. Sa hauteur est donc divisée par 2

I. Généralités sur la statistique

I.3 Présentation des résultats. Histogrammes

Deux histogrammes ne sont comparables que s'ils correspondent au même nombre de résultats.
Pour comparer des histogrammes entre eux, on utilise la notion de fréquence par classe.

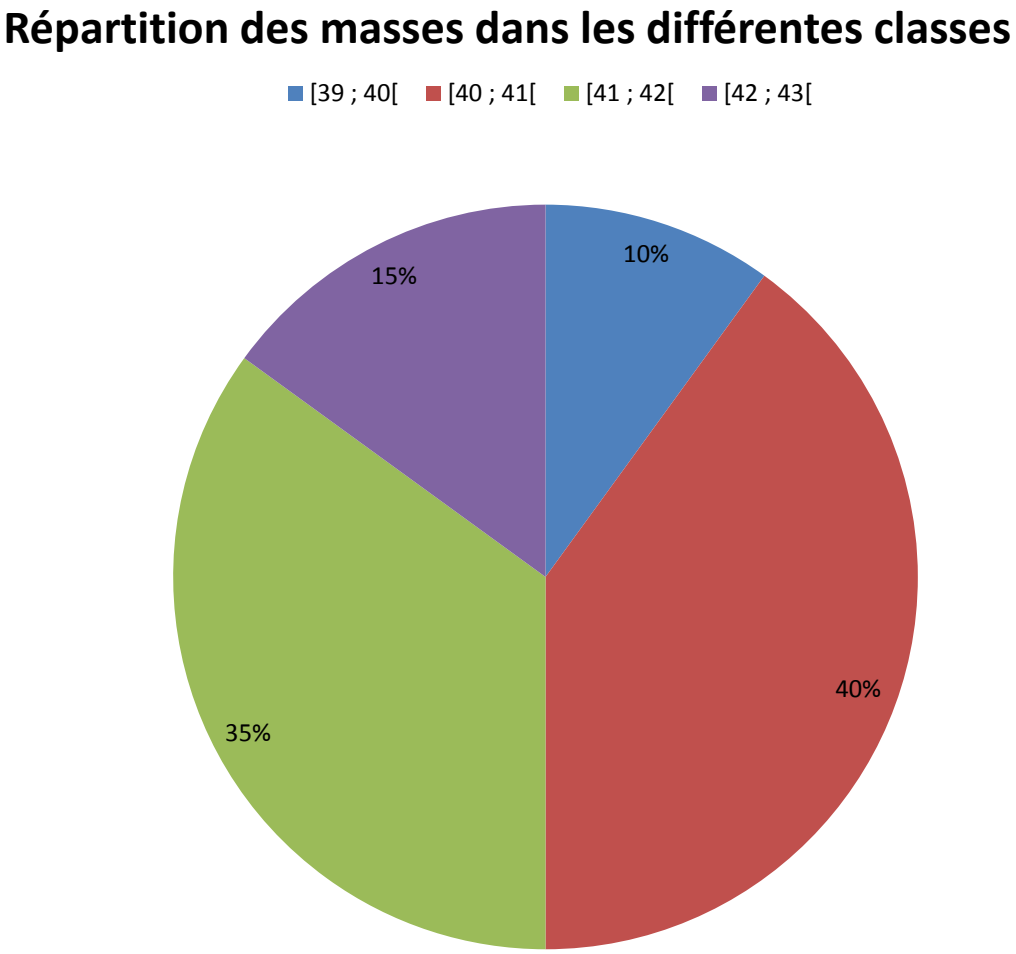
$$f_{AB} = \frac{\text{nombre total de résultats dans la classe } [x_A, x_B[}{\text{nombre total de résultats}}$$

La fréquence peut être donnée en %

On trace ensuite l'histogramme des fréquences.

I. Généralités sur la statistique

On peut représenter les données sous forme de diagramme circulaire en représentant les effectifs (en %) :



II. Séries statistiques simples

Après un certain nombre de mesures, on va chercher à résumer la population par une ou plusieurs caractéristiques (moyenne, médiane, dispersion …)

II.1 Moyenne arithmétique

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

pour des données énumérées

$$\bar{x} = \frac{\sum_{i=1}^{n} n_i x_i}{\sum_{i=1}^{n} n_i} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

pour des données groupées en classes. Dans ce cas xi est le centre de classe ou la moyenne calculée à l'intérieur de la classe.

II. Séries statistiques simples

II.2 Médiane

La médiane est une caractéristique qui partage la population en deux sous-ensembles d'effectifs égaux : le nombre d'observations de part et d'autre de la médiane est donc identique une fois que les observations ont été classées par ordre de grandeur croissante.

La médiane n'est définie en toute rigueur que pour un nombre de mesures impair.
Pour un nombre pair de mesures, on prend la moyenne des deux valeurs centrales.

II. Séries statistiques simples

II.3 Caractéristiques de dispersion

Exemples de deux séries de mesures de même valeur moyenne.

Série 1	95	97	100	103	105
Série 2	50	75	100	125	150

La valeur moyenne seule ne permet pas de caractériser une série de mesures.

L'étendue : $W = x_{\max} - x_{\min}$

L'écart absolu moyen : $e = \dfrac{\sum \left| x_i - \overline{x} \right|}{n}$

II. Séries statistiques simples

II.3 Caractéristiques de dispersion

Variance : $$V = \frac{\sum (x_i - \bar{x})^2}{n} = \sigma_n^2$$

La variance de l'échantillon est la moyenne arithmétique des carrés des écarts des valeurs individuelles à la moyenne générale. La variance permet de comparer entre elles des dispersions de distributions n'ayant pas le même nombre d'observations.

Ecart-type : $$\sigma_n = \sqrt{V} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$ (même unité que la grandeur mesurée)

La série la plus dispersée aura l'écart-type le plus grand.

II. Séries statistiques simples

II.3 Caractéristiques de dispersion

Théorème de Koenig : $V = \sigma_n^2 = \frac{\left(\sum x^2\right)}{n} - \left(\overline{x}\right)^2$

Intervalles de confiance et niveau de confiance :

50% de chances que la vraie valeur soit dans $\left[x - \frac{2\sigma}{3}, x + \frac{2\sigma}{3}\right]$

68% de chances que la vraie valeur soit dans $[x - \sigma, x + \sigma]$

95% de chances que la vraie valeur soit dans $[x - 2\sigma, x + 2\sigma]$

II. Séries statistiques simples

II.3 Caractéristiques de dispersion

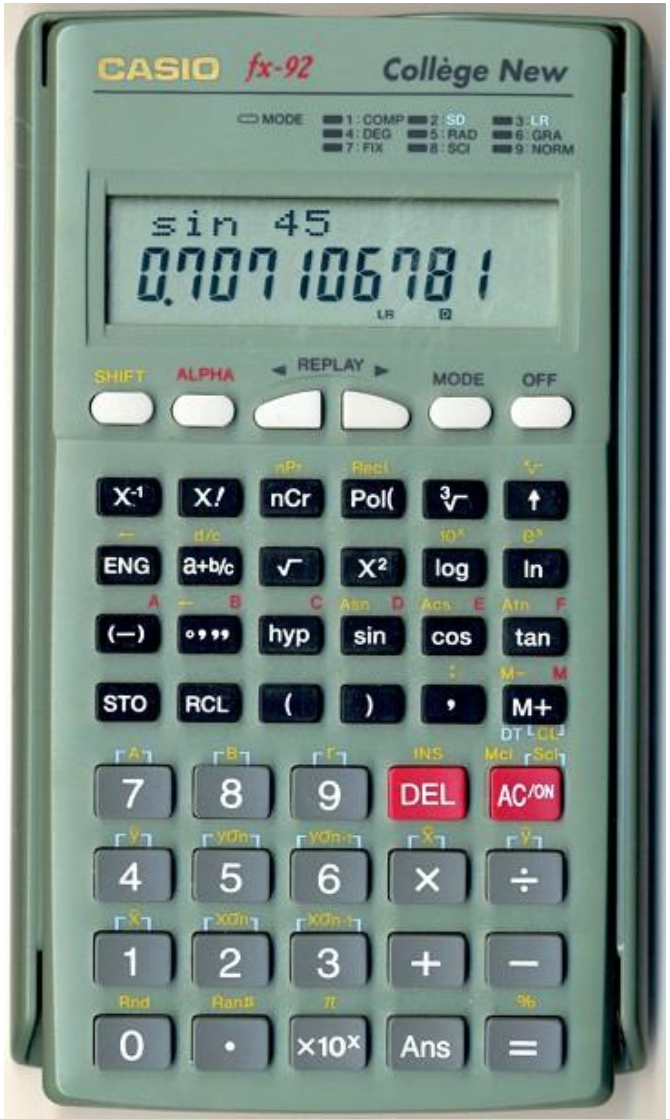
Coefficient de variation : $C_V = \frac{100.\sigma_n}{\bar{x}}$

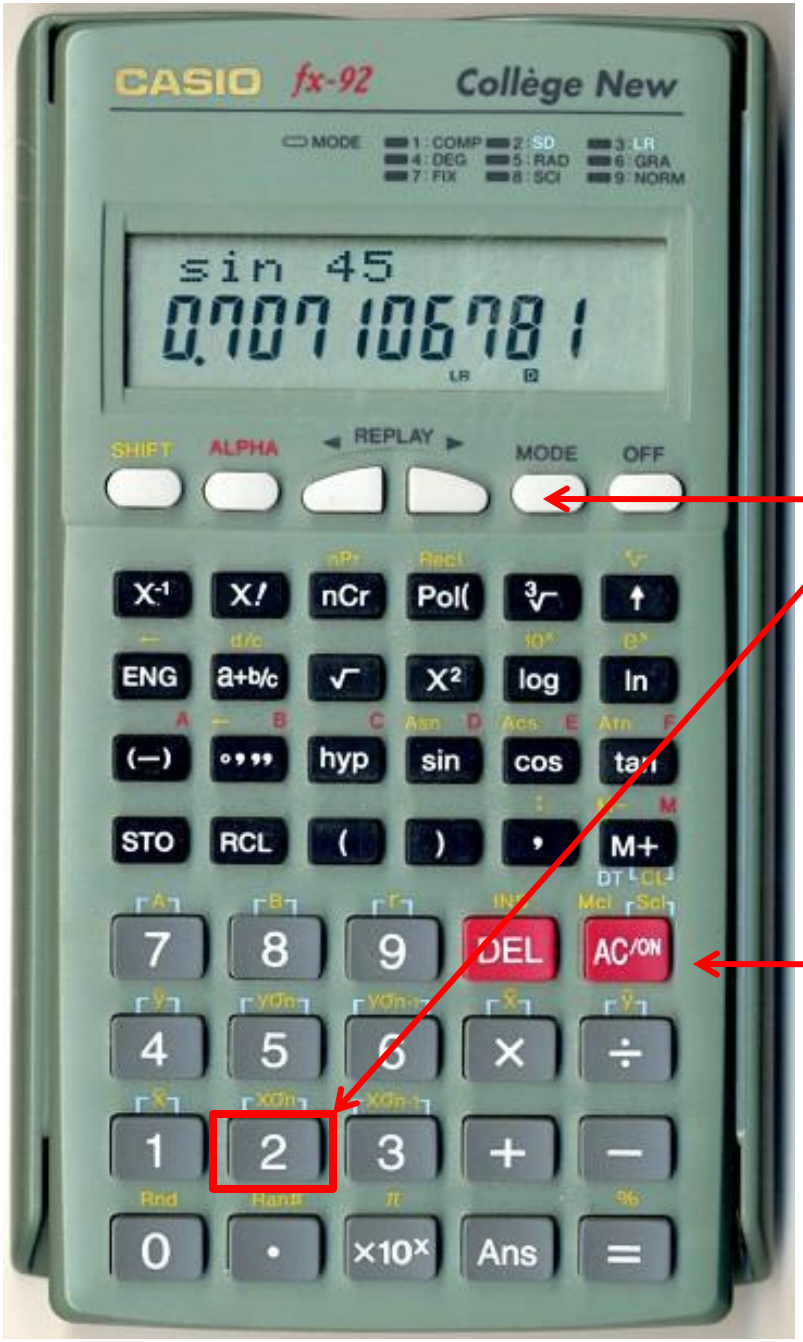
Le coefficient de variation C_v est utilisé dans les cas de séries ayant des ordres de grandeur différents. (L'écart-type du caractère prenant les plus grandes valeurs sera certainement supérieur au second)

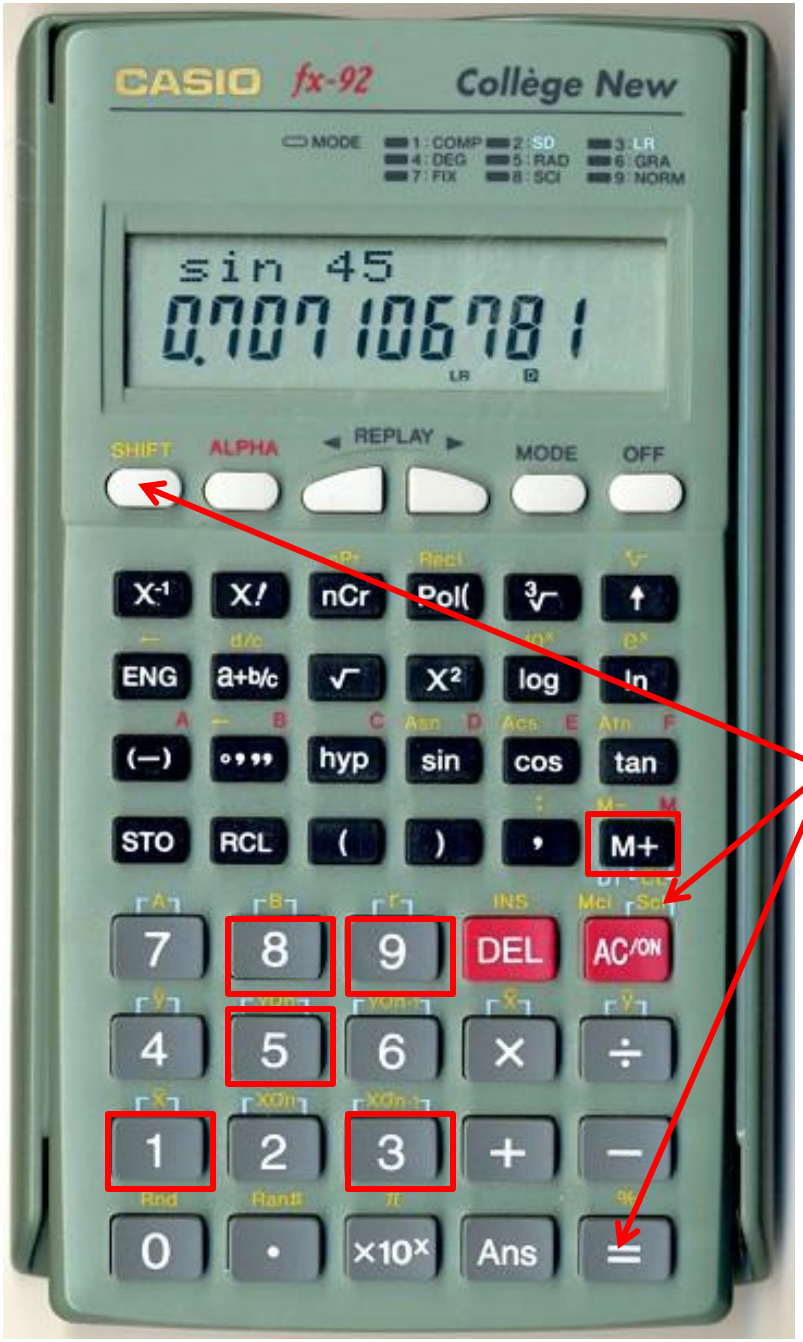
C'est le rapport de l'écart-type sur la moyenne arithmétique exprimé en pourcents (donc sans dimension).

Il relativise l'écart-type par rapport à la moyenne.

Utilisation de la calculatrice en mode statistique



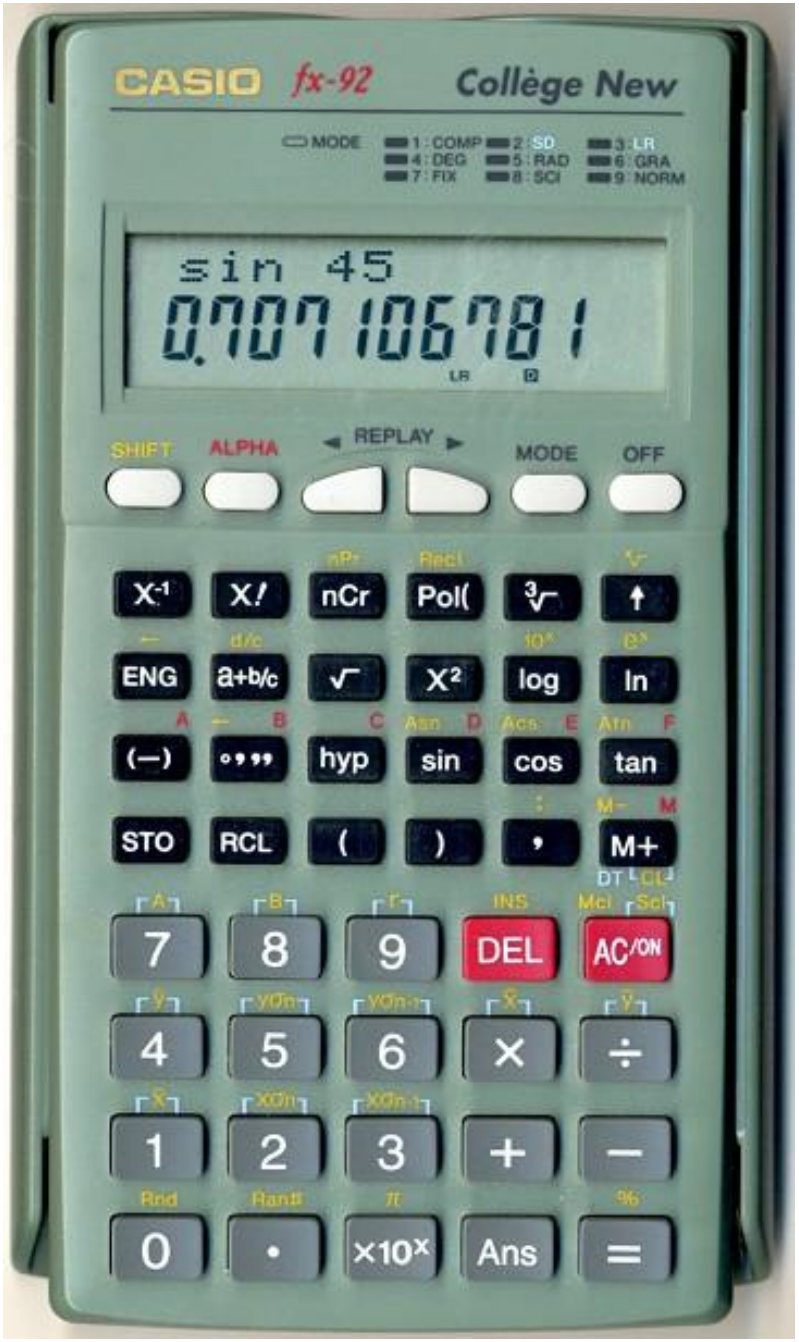

CASIO fx-92 Collège New
MODE
1 : COMP
2 : SD
3 : LR
4 : DEG
5 : RAD
6 : GRA
7 : FIX
8 : SCI
9 : NORM
sin 45
0.707106781
SHIFT
ALPHA
REPLAY
MODE
OFF
ENG
hyp
sin
cos
tan
log
ln
STO
RCL
M+
DEL
AC/ON
Ans
Passer en mode statistique:
« mode 2 »
Allumer la calculatrice



Effacer la mémoire de la calculatrice « shift Scl EXE»

Entrée des valeurs (par exemple 1 3 5 8 et 9)

« 1 M+ 3 M+5 M+ 8 M+ 9 M+



Accès aux différentes caractéristiques

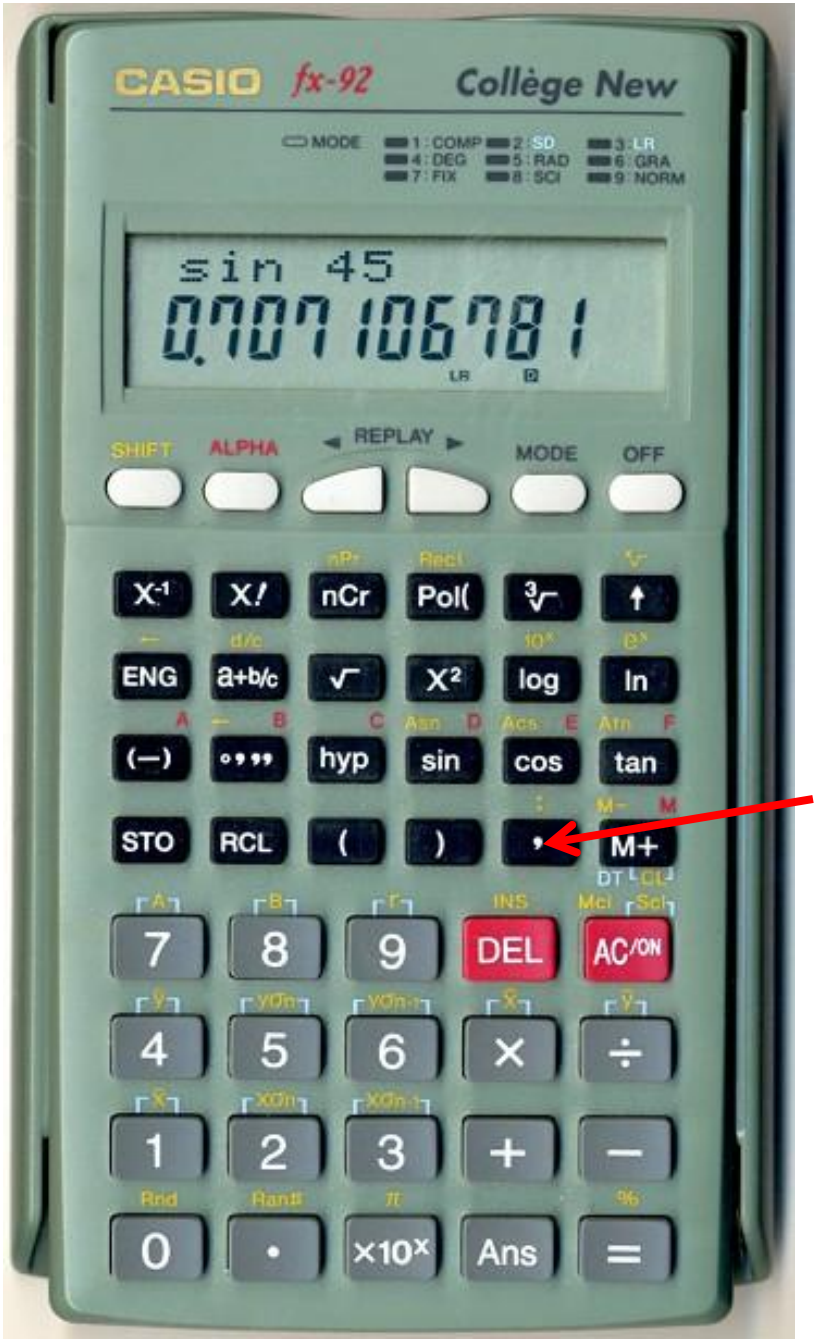
Nombre de mesures n :
« shift 6 »

Moyenne arithmétique :
« shift 1 »

Ecart-type :
« shift 2 »

Somme des carrés :
« shift 4 »

Somme des données :
« shift 5 »



Pour entrer une valeur qui se répète plusieurs fois :

Exemple d'une série de notes d'examen:

8	8	9	10	11	11	11	11	13	13
13	13	13	15	16	17	18	18	19	20

C'est le point-virgule qui permet de réaliser cette opération

8 ; 2 M+ 9 M+ 10 M+ 11 ; 4 M+ etc

Moyenne : 13,4
Ecart-type : 3,6

III. Séries statistiques doubles

On étudie simultanément deux caractères de la population statistique.

- Mise en évidence d'une relation entre ces caractères
- Test de leur degré de dépendance

III.1 Représentation des séries statistiques doubles

a. Distribution

-Liste des valeurs que peuvent prendre les 2 variables observées x_j et y_i.

- Fréquence de ces deux couples dans la population

III. Séries statistiques doubles

III.1 Représentation des séries statistiques doubles

a. Distribution

Exemple : test de la distance de freinage réalisé sur 40 véhicules en fonction de leur vitesse.

y (m) \ x (km/h)	40	60	80	100	120
10	8	4			
30	2	5	4		
50		1	4	2	
70			2	5	
90				3	
110					

Lecture du tableau

y (m) \ x (km/h)	40	60	80	100	120
10	8	4			
30	2	5	4		
50		1	4	2	
70			2	5	
90				3	
110					

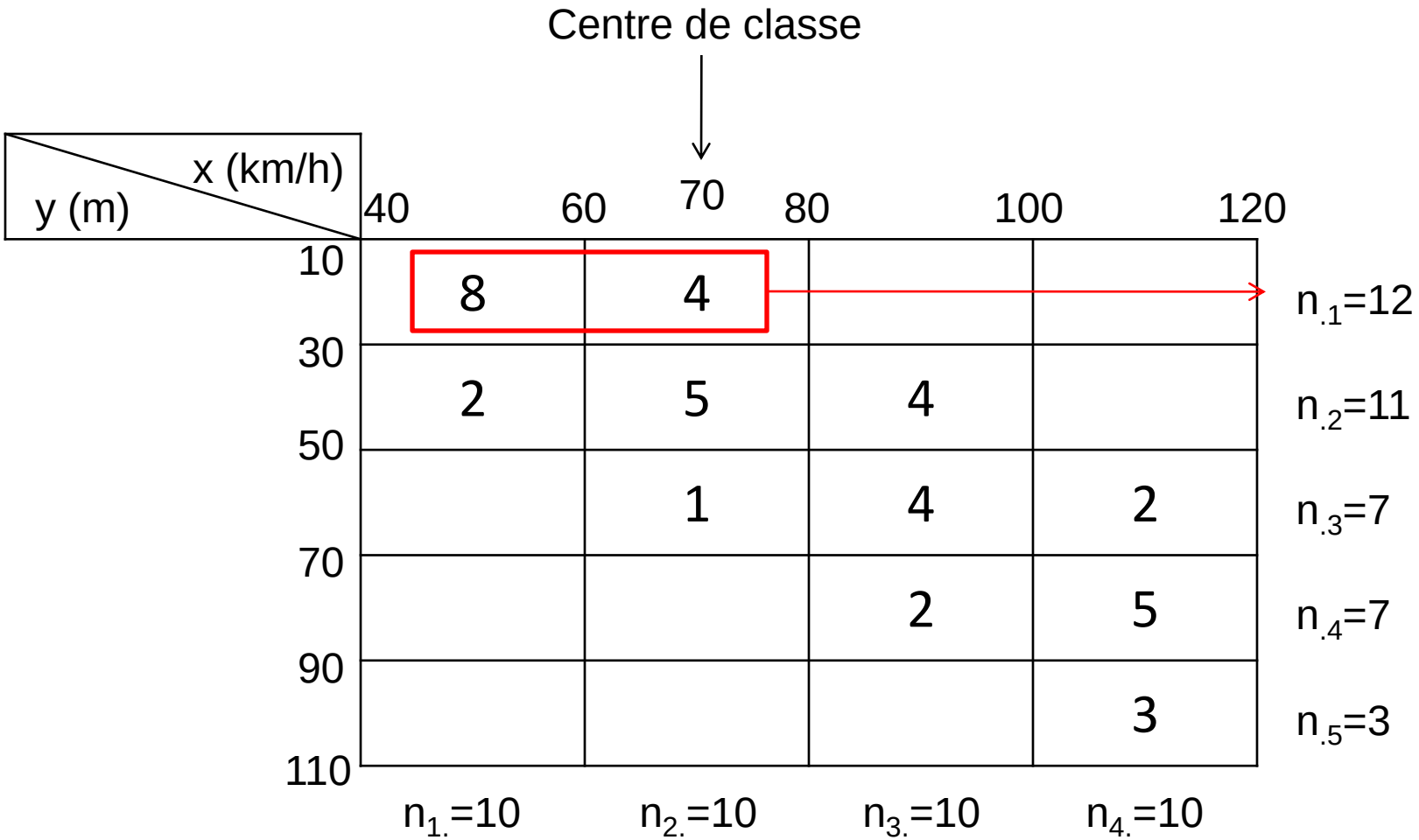
5 véhicules ayant une vitesse comprise entre 60 et 80 km/h se sont arrêtées sur une distance comprise entre 30 et 50 m

b. Distribution marginale

A partir d'une variable à deux dimensions, on peut étudier chaque variable de façon indépendante (calcul de la moyenne, de l'écart-type …)

Une distribution marginale se comporte comme une série statistique simple.

b. Distribution marginale



$$\bar{x} = \frac{10*50+10*70+10*90+10*110}{40} = 80 \text{ km/h} \qquad \sigma_x = 22{,}36 \text{ km/h}$$

$$\bar{y} = \frac{12*20+11*40+7*60+7*80+3*100}{40} = 49 \text{ m} \qquad \sigma_y = 25{,}67 \text{ m}$$

III. Séries statistiques doubles

III.2 Paramètres spécifiques d'une distribution à deux dimensions

Contrairement aux séries simples où l'on peut calculer une moyenne, le « couple moyen » n'a aucune signification pratique.

- Covariance

La variance d'un couple n'existe pas. On parle de covariance.

La covariance est la moyenne arithmétique du produit des écarts aux moyennes arithmétiques respectives de *x* et *y*.

On montre que : $\operatorname{cov}(x,y) = \overline{xy} - \overline{x}\,\overline{y}$

III. Séries statistiques doubles

III.2 Paramètres spécifiques d'une distribution à deux dimensions

- Coefficient de corrélation

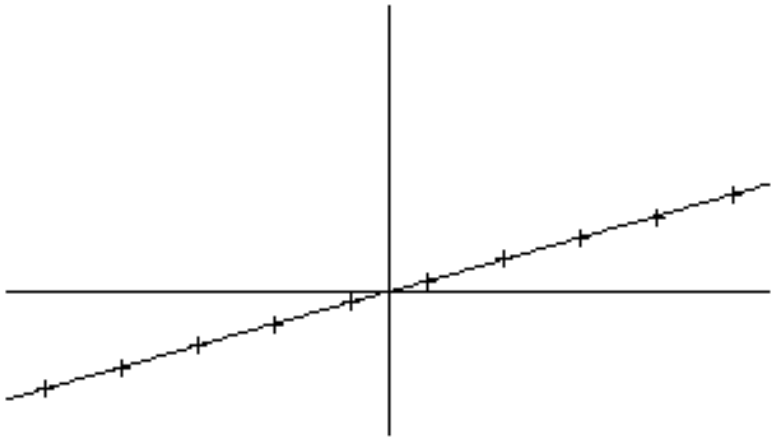
Le coefficient de corrélation permet de mettre en évidence l'existence d'une relation linéaire entre x et y.

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \qquad -1 < r(x,y) < 1$$

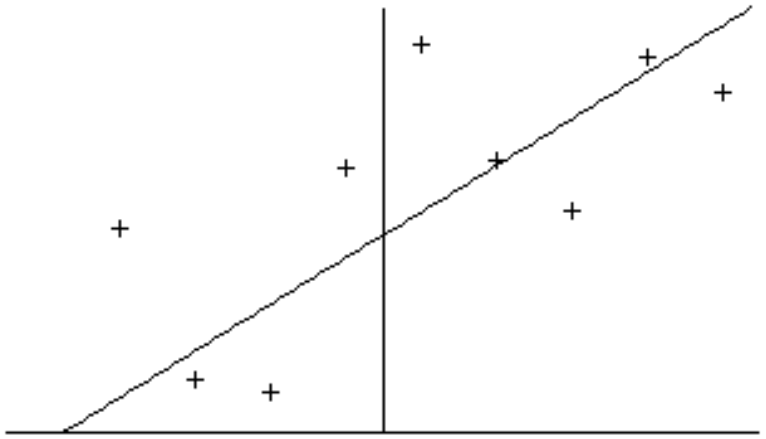
Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte.

Remarque : le coefficient de corrélation est extrêmement sensible à la présence de valeurs aberrantes.

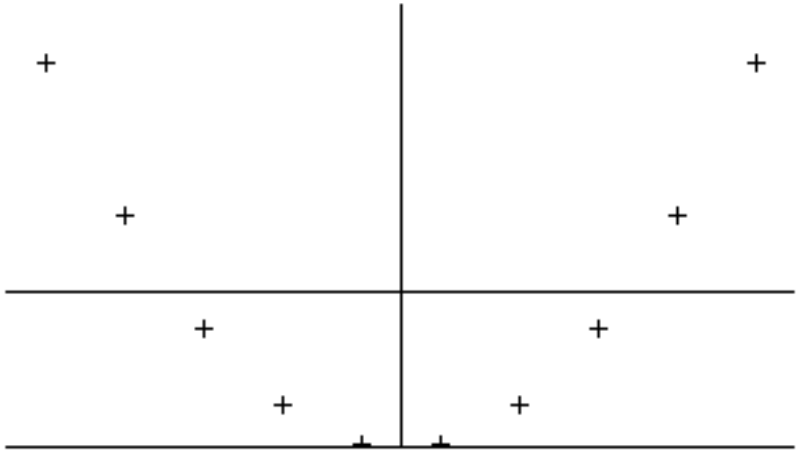
Coefficient de corrélation 1



Coefficient de corrélation 0.77



Coefficient de corrélation 0



Si les deux variables sont totalement indépendantes, alors leur corrélation est égale à 0. La réciproque est cependant fausse, car le coefficient de corrélation indique uniquement une dépendance *linéaire*

y (m) \ x (km/h)	40	60	70	80	100	120
10	8	4				
30	2	5		4		
50		1		4	2	
70				2	5	
90					3	
110						

$\bar{x} = 80\ \text{km/h}$ $\sigma_x = 22{,}36\ \text{km/h}$ $\overline{xy} = \dfrac{8*50*20 + 4*70*20 + \ldots}{40} 4410$

$\bar{y} = 49\ \text{m}$ $\sigma_y = 25{,}67\ \text{m}$

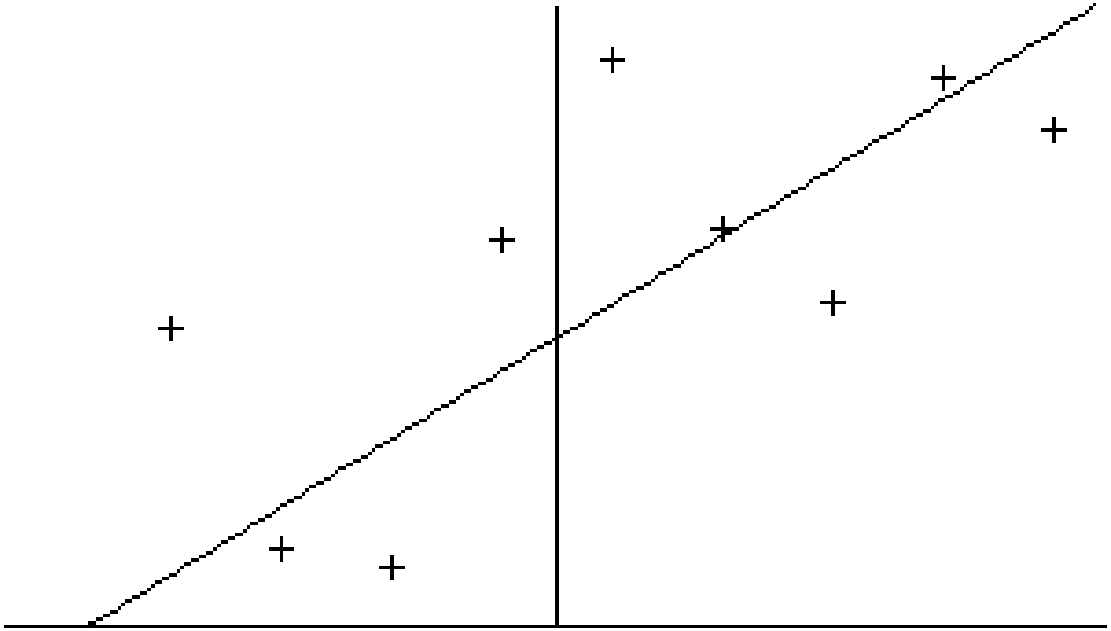
$$\text{cov}(x, y) = \overline{xy} - \bar{x}\bar{y} = 490$$

$$r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = 0{,}85$$

III. Séries statistiques doubles

III.3 Ajustement linéaire. Méthode des moindres carrés.

Si on porte sur un graphique les points représentatifs des x_i et y_i, on obtient un nuage de points. On peut déterminer une droite qui « résume » l'ensemble des points.



III. Séries statistiques doubles

III.3 Ajustement linéaire

a. Ajustement graphique

On trace au jugé une droite D passant par le plus près possible des points du nuage de points, **en s'efforçant d'équilibrer** le nombre de points situés au dessus et au dessous de la droite D.

b. Méthode des moindres carrés

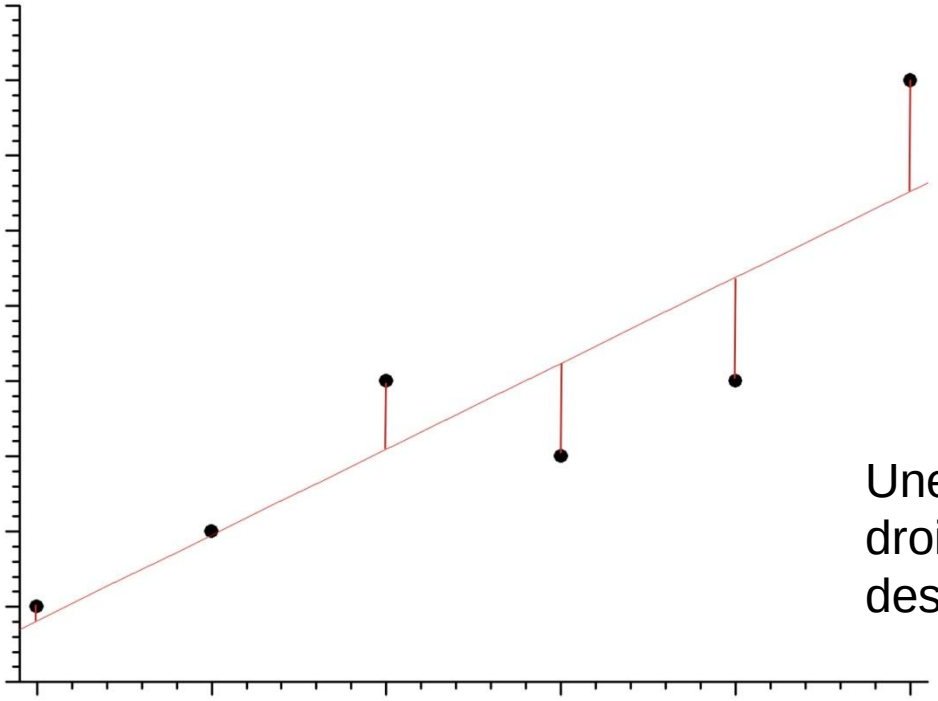
La **méthode des moindres carrés**, indépendamment élaborée par Legendre en 1805 et Gauss en 1809, permet de comparer des données expérimentales, généralement entachées d'erreurs de mesure à un modèle mathématique censé décrire ces données

On cherche une droite telle que la somme de ses « distances » aux différents points représentant les données soit minimale.

III. Séries statistiques doubles

III.3 Ajustement linéaire

b. Méthode des moindres carrés



Une seule droite (appelée meilleure droite) permet de minimiser la somme des écarts à la meilleure droite

III. Séries statistiques doubles

III.3 Ajustement linéaire

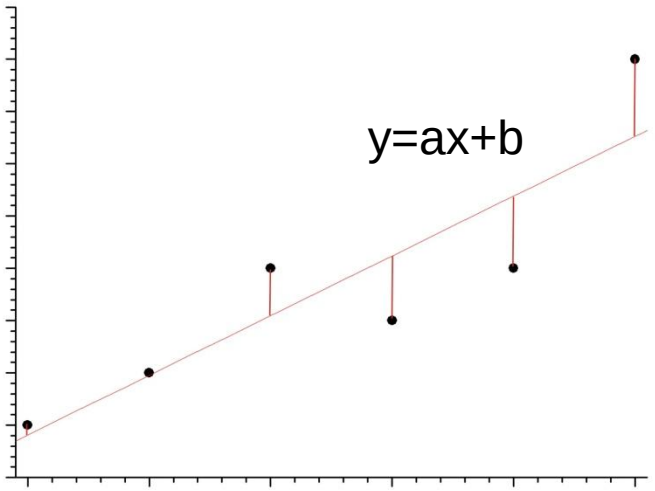
b. Méthode des moindres carrés

Droite de régression de y en x.

$$y = ax + b$$

$$a = \frac{\mathrm{cov}(x, y)}{V_x}$$

$$b = \overline{y} - a\overline{x}$$



III. Séries statistiques doubles

III.3 Ajustement linéaire

c. Méthode de Mayer

On partage le nuage de points en deux nuages de points de nombres équivalents. On calcule alors le point moyen de chaque nuage qu'on appelle G1 et G2. La droite (G1G2) est la droite de Mayer.
C'est une bonne approximation, si le nuage de points est allongé.

année		01	02	03	04	05	06
CA	M€	15	20	25	25	30	50

$\overline{x} = 02$ $\quad$ $\overline{x} = 05$

$\overline{y} = 21{,}6$ $\quad$ $\overline{y} = 35$