VI. Statistique descriptive.

1. Avant - propos : le signe sommatoire.

Soient $x_1, x_2, ...x_n$: n réels $\qquad x_1 + x_2 + ...+ x_n = \sum_{i=1}^{n} x_i$

Remarquons : $\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$

Propriétés.

1. $\sum_{i=1}^{n} x_i = \sum_{i=1}^{p} x_i + \sum_{i=p+1}^{n} x_i \qquad (p < n)$
2. $\sum_{i=1}^{n} kx_i = k \sum_{i=1}^{n} x_i$
3. $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

Exercices.

1. Montrer $\sum_{i=1}^{n} (x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$
2. Décomposer $\sum_{i=1}^{i=n} (x_i - y_i)^3$
3. Montrer : $\sum_{i=1}^{n} (x_i + a) = \left(\sum_{i=1}^{n} x_i\right) + na$
4. Démontrer (par récurrence)
   a) $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$
   b) $\sum_{i=1}^{n} i^3 = \left(\sum_{i=1}^{n} i\right)^2$

2. Statistique : rappel des éléments vus en quatrième.

2.1 Généralités

Statistique est un terme souvent rencontré dans le langage courant actuellement: on en trouve dans les journaux, à la T.V, dans le domaine scientifique, politique, économique ....
Dans ce chapitre, nous envisagerons les règles de base permettant de rassembler les données, calculer les valeurs centrales et de dispersion mais aussi nous chercherons à avoir un regard critique face aux chiffres. En effet, il faut se rendre compte que les entreprises peuvent utiliser les données à leur avantage. A partir de mêmes données de départ, on peut mettre en évidence des choses tout à fait différentes.
Exemple caricatural : dans une première classe, sur 25 élèves, il y en a 7 qui ont plus de 85% et dans une seconde classe, tous les élèves ont moins de 85%. Si on s'arrête à ces chiffres, on pourrait croire que la première classe est plus forte que la deuxième. Si on regarde un peu plus loin, tous les élèves de la deuxième classe ont entre 65 et 75%, alors que les 18 restant dans la première ont moins de 65%. Alors, quelle est la meilleure des deux classes ? On remarque que suivant deux points de vue différents, on peut tirer des conclusions très différentes.
Un autre exemple pourrait être l'interprétation des mêmes chiffres d'une enquête sur les effets de la cigarette par des médecins ou par les fabriquants de cigarettes.
Enfin, avant de passer à la formalisation, prenons quelques exemples où la démarche statistique sera employée.
Exemple 1
Une usine fabrique 500 000 ampoules électriques par jour. On désire connaître la proportion d'ampoules défectueuses dans la production. Comment s'y prendre ? Tester toutes les ampoules ?
Solution exacte : mettre chacune de 500 000 ampoules dans un soquet et les tester.
Solution statistique : choisir 1000 ampoules parmi les 500 000 (échantillon)
Les tester pour obtenir la solution exacte pour ces 1000 ampoules
Considérer cette solution comme estimateur de la proportion exacte.

Exemple 2 tiré de la pharmacologie:
Un médecin désire tester l'effet d'un médicament sur le rythme cardiaque; 50 malades cardiaques sont choisis et traités avec ce médicament. On note pour chacun d'eux l'augmentation du taux de pulsations. Ensuite, après avoir examiné ces résultats, le médecin infère que ce médicament aura les mêmes effets sur tous les futurs patients.
Exemple 3 Sondages d'opinion.
Pour prédire le résultat d'élections en France, l'institut choisit 1200 personnes sur 30 000 000 votants. Le résultat de cette prédiction s'avère souvent vérifié.

A travers ces exemples, nous voyons donc que le travail du statisticien comportera trois parties
- L'échantillonnage et la collecte des données (les résultats sont aléatoires)
- L'analyse des données.
- L'inférence à propos d'un plus grand ensemble de données: la population.

Remarques sur l'échantillonnage:
1. Cet échantillonnage est souvent indispensable pour différentes raisons:
   - coût trop élevé
   - facteur temps (ex : cela prendrait trop de temps de tester 500 000 ampoules)
   - taille de la population (ex : 500 000 ampoules, 30 000 000 votants)
   - inaccessibilité de la population.
   - nature destructive de l'observation (tester la durée de vie d'ampoules: on ne peut les tester toutes pour voir combien de temps elles résistent)
2. L'échantillon n'est pas toujours représentatif de la population
   - On peut parfois tomber dans une catégorie spécifique (ex : on va au hasard dans la rue pour faire une étude de la taille des gens, et on croise une équipe de basketteurs)
   - Le cas le plus fréquent: certains éléments de la population n'ont aucune chance de figurer dans l'échantillon. Un exemple typique de cette situation est donné par les élections américaines de 1936 entre Roosevelt et Landon. "Literary digest" avait interrogé plusieurs (3) millions de personnes et avait prédit une victoire de Landon avec un écart record. Or le résultat fut exactement inverse. Cette erreur était due au fait que l'échantillon avait été choisi dans les annuaires téléphoniques. Or, au sortir de la crise, peu nombreux étaient les démocrates qui pouvaient se permettre un tel luxe.

Ces types d'erreur sont très difficiles à éviter. Un échantillon 10 fois plus grand ne donnera pas nécessairement des résultats 10 fois meilleurs.
De même à partir de résultats statistiques, il faudra se méfier: si dans un jeu on a 95% de chances de gagner, il restera toujours 5% de chances de perdre et on ne peut donc prévoir à l'avance si on va gagner ou perdre !

2.2 Démarche statistique : Formalisation.

La statistique recueille et étudie des observations sur des ensembles de même nature (personnes, animaux, objets...)
L'ensemble étudié est appelé ***population***
Souvent l'étude n'est faite que sur une partie de la population appelée ***échantillon.***
L'étude se fait sur un trait de la population appelé ***caractère.***
On rencontre deux types de caractères
- Quantitatif (ex : mesure de températures, taille d'une personne, âge ....)
- Qualitatif (ex : profession, couleur des cheveux, moyen de locomotion...)

Parmi les caractères quantitatifs, deux situations peuvent se présenter:
- caractères discrets : nombres entiers (ex: nombre d'enfants dans une famille, nombre de buts d'une rencontre de football,...)
- caractères continus : toutes les valeurs d'un intervalle sont possibles (ex : taille ou poids d'une personne...)

3. Analyse de données : variable statistique à caractère discret

3.1 Présentations numériques.

Prenons comme exemple la population des rencontres de football d'un week-end dont les résultats sont publiés dans les pages sportives d'un journal donné. Comme variable statistique, nous prendrons le nombre total de buts marqués au cours de chacune des rencontres. Nous constituons ainsi un échantillon de 50 rencontres, ce qui nous donne les nombres suivants:

2	0	2	8	2	2	4	2	0	4	0	5	2	2	5	2	0	3	3	3	0	2	2	1	3
2	4	0	2	7	8	1	1	5	3	6	2	3	1	2	1	0	4	5	2	4	3	1	1	6

Ce tableau ainsi présenté appelé ***tableau brut*** n'est pas très parlant. Nous allons donc ordonner ces résultats pour obtenir le ***tableau ordonné***.

0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	5	5	5	5	6	6	7	8	8

Cette écriture reste quand même lourde, et il paraît plus intéressant de n'écrire qu'une seule fois les résultats se répétant en retenant le nombre de fois qu'ils apparaissent. Nous allons ainsi obtenir le ***tableau groupé***.

Nombre de buts	Nombre de rencontres.
0	7
1	7
2	15
3	7
4	5
5	4
6	2
7	1
8	2

Dans ce tableau, les différentes valeurs du caractère sont désignées par x_i (modalités)
L'effectif des membres présentant le caractère x_i est désigné par r_i (Répétitions)

$n = \sum_{i=1}^{p} r_i$ est la somme des effectifs des différentes modalités et vaut l'effectif total de l'échantillon.

p est le nombre de modalités distinctes.
Dans notre exemple, les x_i sont les valeurs 0, 1, 2, 3 .....
les r_i sont les valeurs 7, 7, 15, 7, ...
le nombre de modalités = 9
Afin d'avoir une idée plus précise de la proportion dans laquelle un résultat apparaît, nous pouvons calculer sa ***fréquence***.(f_i)
exemple : 0, 1, et 3 ont une fréquence égale à 7/50
2 a pour fréquence 15/50……..
Nous pouvons aussi nous demander combien de matches ont eu un nombre de buts inférieur ou égal à la valeur considérée et obtenir ainsi les ***effectifs cumulés***.
exemple : 2 a pour effectif cumulé 29
3 a pour effectif cumulé 36
De là on déduira facilement les ***fréquences cumulées*** (F_i): proportions des résultats inférieurs ou égaux à une valeur donnée.
Toutes ces valeurs sont rassemblées dans le ***tableau recensé.***

Modalités (Nbre de buts) x_i	Effectifs (Nbre de rencontres) r_i	fréquences f_i	effectifs cumulés R_i	fréquences cumulées F_i
0	7	7/50	7	7/50
1	7	7/50	14	14/50
2	15	15/50	29	29/50
3	7	7/50	36	36/50
4	5	5/50	41	41/50
5	4	4/50	45	45/50
6	2	2/50	47	47/50
7	1	1/50	48	48/50
8	2	2/50	50	50/50

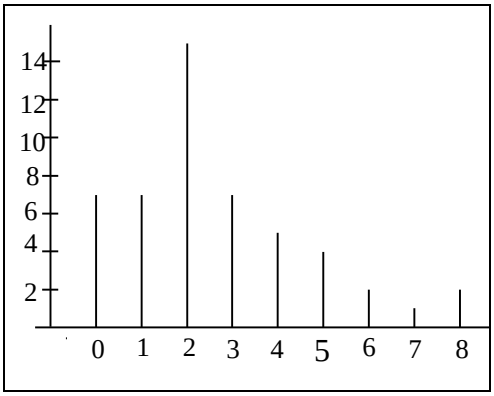
Mais cette présentation ne nous montre pas encore facilement les résultats: les représentations graphiques seront plus explicites.

3.2 Représentations graphiques.

3.2.1 Le diagramme en bâtonnets.

On porte en abscisse les différents résultats et on trace sur ces valeurs des "bâtonnets" de hauteurs proportionnelles aux effectifs de chaque modalité.

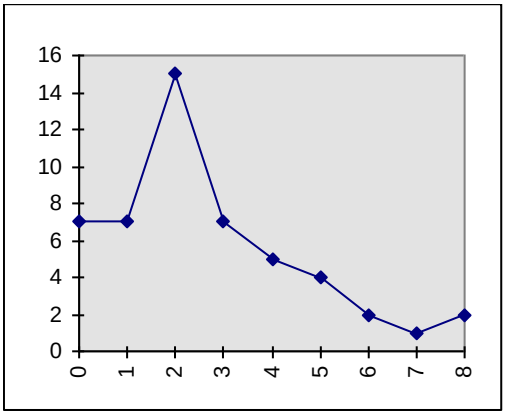
On obtiendrait un diagramme équivalent à l'échelle près si on portait en ordonnée les fréquences respectives des différentes modalités.



3.2.2 Le polygone des effectifs.

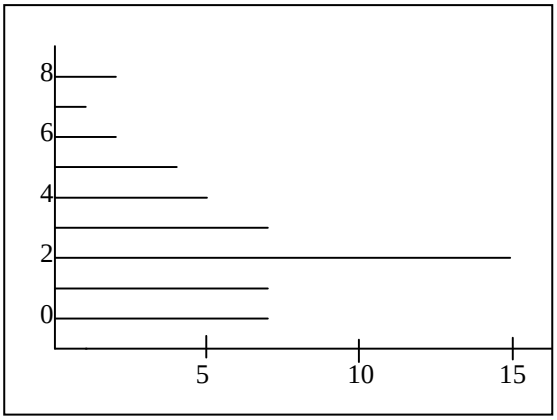
Ce diagramme s'obtient de manière semblable au précédent.
On porte en abscisse les différents résultats et en ordonnée, les effectifs correspondants et on relie les points obtenus.

A nouveau, on obtient un diagramme équivalent à l'échelle près en portant en ordonnée les fréquences au lieu des effectifs.



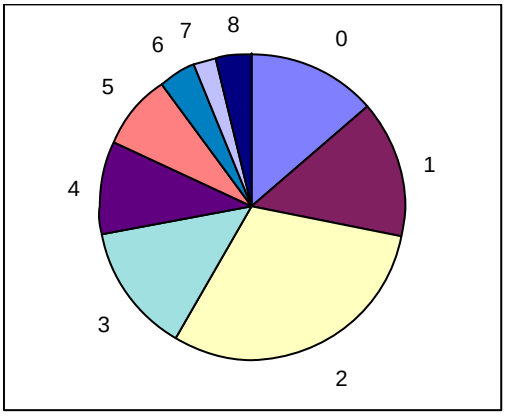
3.2.3 Le diagramme en pyramide.

Ce diagramme est semblable au diagramme en bâtonnets : on inverse simplement l'abscisse et l'ordonnée.



3.2.4 Le diagramme circulaire.

Appelé parfois diagramme en camembert, ce dernier est un diagramme d'aire : la surface de chaque secteur circulaire est proportionnelle à la répétition de la modalité qu'il représente.



3.3 Valeurs centrales.

En observant les tableaux et graphiques, nous pouvons nous poser différentes questions quant au résultat global.

- Est-ce la moyenne arithmétique des valeurs (c-à-d la somme des résultats divisée par le nombre de résultats)
- ou la valeur qui apparaît le plus souvent ?
- ou encore la valeur telle qu'il y ait autant de résultats inférieurs à celui-ci que de résultats supérieurs.

En fait, ces trois valeurs ont chacune leur importance.

1. $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} r_i x_i$ est la ***moyenne*** de l'échantillon (p est le nombre de résultats distincts).
   Dans notre cas $\bar{x}$ = 2,66
2. Le ***mode*** d'une série statistique est la valeur qui apparaît le plus souvent.
   Dans notre exemple le mode est la valeur 2
3. La ***médiane*** d'une série statistique est la valeur telle qu'il y ait au maximum 50 % des résultats strictement inférieurs à celle-ci et au maximum 50 % de résultats qui lui sont strictement supérieurs
   Dans notre cas, la médiane vaut 2. En effet, $\frac{14}{50}$ des résultats sont strictement inférieurs à 2 et $\frac{21}{50}$ des résultats sont strictement supérieurs à 2
   Parfois deux valeurs pourraient être considérées comme médiane : dans ce cas, on prendra pour médiane, la moyenne entre ces deux valeurs.

3.4 Indices de dispersion.

Un dernier point d'observation est encore fort utile : la répartition des résultats. Sont-ils groupés autour de la valeur centrale ou au contraire fortement dispersés ?
Quatre éléments nous donnent des réponses à ces questions.

1. *L'étendue* d'un tableau statistique est la différence entre la plus grande valeur et la plus petite.
   Notre exemple : l'étendue = 8
2. *Les écarts à la moyenne* des différents résultats sont les différences en valeur absolue entre la moyenne et ces valeurs = $|x_i - \bar{x}|$

x_i	0	1	2	3	4	5	6	7	8
Ecarts à la moyenne	2.66	1.66	0.66	0.34	1.34	2.34	3.34	4.34	5.34

3. La *variance* d'une série statistique, notée σ^2 est la moyenne arithmétique des carrés des écarts à la moyenne.

$$s^2 = \frac{1}{n}\sum_{i=1}^{p} r_i (x_i - \bar{x})^2$$ Dans notre exemple s^2 = 4.1044.

4. Enfin *l'écart type*, noté s est la racine carrée de la variance.
Dans notre exemple : s = 2.02593

Remarque : parfois, les machines utilisent $\frac{1}{n-1}$ au lieu de $\frac{1}{n}$. (Parfois, on trouvera une touche s_{n-1}). Il suffira alors de multiplier les résultats par $\frac{n-1}{n}$

4. Groupement des données en classes.

Parfois, le nombre de résultats différents devient trop important. C'est le cas pour une série statistique à valeurs discrètes dont le nombre de valeurs distinctes est grand mais c'est surtout le cas pour une série statistique à valeurs continues. Nous allons alors diviser les résultats en classes, et ensuite faire le même travail que dans le cas précédent. Chaque classe est désignée par sa valeur centrale.
Quelques notations préalables.

- *L'étendue d'une classe* est la différence entre ses extrémités.
- La *valeur centrale* d'une classe sera notée x_i et vaut le centre de l'intervalle ($i^{ème}$ classe)
- Les *répétitions* ou effectifs des classes : nombre d'éléments de cette classe noté r_i pour la $i^{ème}$ classe.
- Le nombre de classes noté p
- *L'effectif de la série statistique* qui vaut la somme des répétitions noté n.

- La *<u>fréquence</u>* d'une classe : $f_i = \frac{r_i}{n}$
- La *<u>fréquence cumulée</u>* de chaque classe qui est la somme des fréquences de cette classe et de celles qui la précèdent. $F_s = f_1 + f_2 + ...+f_s = \sum_{i=1}^{s} f_i$

<u>Propriété</u> : Dans une série statistique d'effectif total n comprenant p classes d'effectifs respectifs r_1 , r_2 , ...r_p (n, p $\in$ N) $\quad \sum_{i=1}^{p} f_i = 1$

En effet $\sum_{i=1}^{p} f_i = f_1 + f_2 + ...+f_p = \frac{r_1}{n} + \frac{r_2}{n} + \ldots\ldots \frac{r_p}{n} = \frac{r_1 + r_2 + \ldots\ldots + r_p}{n} = \frac{n}{n}$

4.1 Tableau recensé d'une série statistique groupée par classes.

On a mesuré en cm la taille de 54 enfants; Cela a donné les résultats suivants.

127	129	126	132	125	133	131	128	133	120	135	127	125	112
134	125	136	132	130	123	121	129	133	127	133	135	131	115
134	130	128	132	127	126	141	138	118	136	139	138	138	122
146	134	143	128	142	133	136	131	132	124	127	134		

Comme dans le paragraphe précédent, nous allons ordonner ces résultats, mais surtout les grouper en classes et calculer ensuite les répétitions des classes, leurs fréquences, leurs fréquences cumulées. L'ensemble de ces valeurs est repris dans le *<u>tableau recensé</u>* qui suit.

classes	valeurs centrales x_i	Répétitions r_i	Rép. cumulées R_i	Fréquences f_i	Fréq. cumulées. F_i
[112,118[	115	2	2	2/54 ≅ 0.037	2/54 ≅ 0.037
[118,124[	121	5	7	5/54 ≅ 0.093	7/54 ≅ 0.129
[124,130[	127	16	23	16/54 ≅ 0.0.296	23/54 ≅ 0.426
[130,136[	133	20	43	20/54 ≅ 0.37	43/54 ≅ 0.796
[136,142[	139	8	51	8/54 ≅ 0.148	51/54 ≅ 0.944
[142,146]	144	3	54	3/54 ≅ 0.056	54/54 = 1

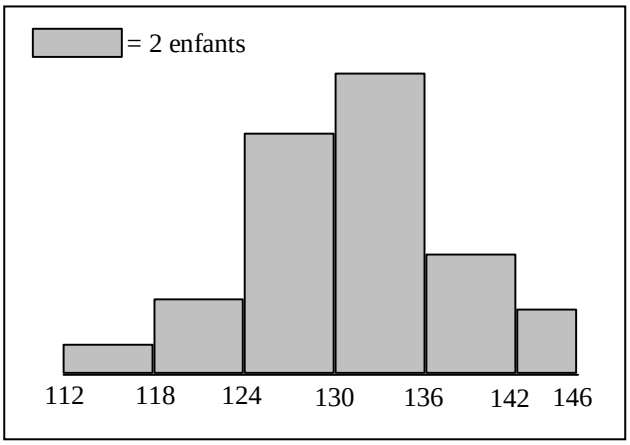
Il n'y a pas vraiment de méthodes pour déterminer le nombre de classes.

Certains choisissent de prendre $(2.5\sqrt[4]{n}\,)$classes, d'autres $(\log_2 n)$ classes.
Le plus souvent, on choisit un nombre de classes compris entre 5 et 20.

4.2 Représentations graphiques d'une série statistique groupée par classes.

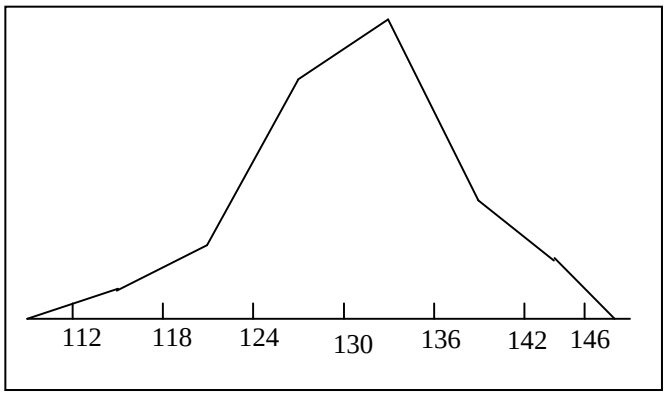
4.2.1 Histogramme.

On remplace le diagramme en bâtonnets par l'histogramme qui est un diagramme d'aires. En abscisse on porte les différentes classes, et sur chaque classe, on construit un rectangle dont la surface est proportionnelle au nombre d'éléments de la classe. Les hauteurs des rectangles ne seront donc pas proportionnelles aux effectifs des classes lorsque celles-ci ne sont pas toutes de même largeur. (c'est le cas de la dernière classe ici) Le graphique obtenu en remplaçant les effectifs par les fréquences est le même à l'échelle près.



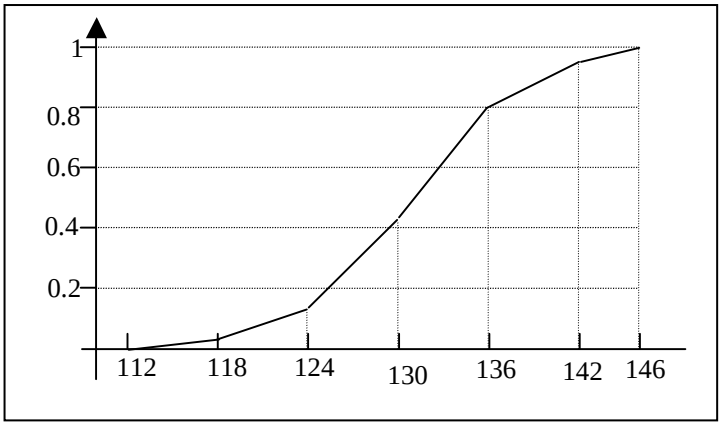
4.2.2 Polygone des effectifs.

Comme dans le cas de statistiques discrètes, on peut obtenir le polygone des effectifs à partir de l'histogramme en reliant les centres des côtés supérieurs des rectangles précédents.
Remarque : On convient parfois de compléter l'histogramme par 2 classes de même amplitude à effectif nul, l'une à gauche et l'autre à droite des classes extrêmes et on joint par des segments de droite les milieux de toutes les bases supérieures des rectangles ainsi obtenus.



4.2.3 Diagramme cumulatif.

On ne connaît les fréquences cumulées que des extrémités des classes. Entre ces points, on ne connaît pas les fréquences cumulées exactes. C'est pourquoi, après avoir placé ces points, on les rejoint par des segments de droites qui donnent des valeurs approchées des fréquences cumulées des valeurs intérieures aux classes.
A nouveau, on obtient un diagramme équivalent à partir des effectifs cumulés plutôt qu'à partir des fréquences cumulées.

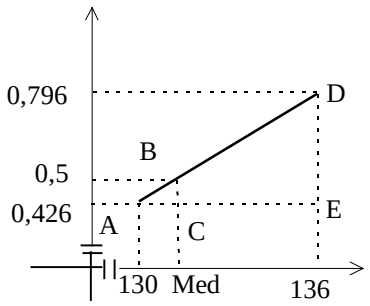


4.3 Valeurs centrales d'une série statistique groupée en classes.

Le mode est la classe dont la fréquence est maximale : ici la classe [130,136] ou la classe de 133
La moyenne est simplement la moyenne arithmétique de la série en ramenant toutes les valeurs d'une même classe à la valeur centrale de celle-ci c.-à-d. $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} r_i x_i$ où n est l'effectif total de la série, p le nombre de classes , r_i la répétition de la $i^{ème}$ classe et x_i la valeur centrale de celle-ci.

Dans notre exemple : $\bar{x} = \frac{1}{54}(2 . 115 + 5 . 121 + 16 . 127 + 20 . 133 + 8 . 139 + 3 . 144) = 130.9444$

la médiane est la valeur telle que 50% des effectifs sont supérieurs à celle-ci et 50% lui sont inférieurs c.-à-d. la valeur dont la fréquence cumulée vaut 0.5
Dans notre exemple, elle ne peut être déterminée directement mais sera approximée à partir du diagramme cumulatif.
En effet en reprenant une partie de celui-ci, nous obtenons le graphique ci-contre



Les triangles ABC et ADE sont semblables et donc $\frac{\overline{AC}}{\overline{AE}} = \frac{\overline{BC}}{\overline{DE}} \Leftrightarrow$

$\frac{\overline{AC}}{136-130} = \frac{0.5-0.426}{0.796-0.426} \Leftrightarrow \frac{\overline{AC}}{6} = \frac{0.074}{0.37} \Leftrightarrow \frac{\overline{AC}}{6} = 0.2 \Leftrightarrow \overline{AC} = 1.2$ et la médiane vaut donc 130 + 1.2 = 131.2

Remarque : La droite x = Med partage l'histogramme en 2 parties de surfaces égales.
On peut de même déterminer les *premier* et *troisième quartiles:* : valeurs dont les fréquences cumulées valent respectivement 0.25 et 0.75, c.-à-d. valeurs telles que 25% ou 75% de l'effectif leur sont inférieurs.
On procédera comme pour la médiane pour les déterminer.
L'intervalle interquartile = la différence entre le troisième et le premier quartile (q_3 - $q_{1)}$; dans cet intervalle se situe 50% de l'effectif total.
Et de même, on peut montrer que les droites $x = q_1$, $x = q_2$ et $x = q_3$ partagent l'histogramme en 4 parties égales.

4.4 Indices de dispersion d'une série statistique groupée en classes.

Comme dans le cas d'un tableau non groupé en classes, 3 indices de dispersion seront employés : l'étendue, la variance et l'écart type.

L'étendue d'un tableau est la différence entre les deux valeurs extrêmes.
Dans notre exemple : l'étendue vaut : 146 - 112 = 34
Mais cette valeur ne donne pas une bonne indication sur la dispersion des valeurs du caractère.

La variance est la moyenne des carrés des écarts à la moyenne c.-à-d. : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{p} r_i (x_i - \overline{x})^2$

Dans notre exemple : $s^2 = 43.83036554$
Enfin *l'écart type* est comme précédemment la racine carrée de la variance. Ceci nous permet de retrouver une cohérence dans les unités : la variance dans notre exemple est exprimée en cm^2 tandis que l ' écart - type sera exprimé en cm.
Dans notre exemple s = 6.62045
Remarque
L'écart-type prendra tout son sens lors de l'étude de certaines distributions dites "normales". On constatera en effet que pour ces distributions :

- la surface de l'histogramme comprise entre la moyenne moins un écart-type et la moyenne plus un écart-type vaut 68% de la surface totale de l'histogramme.
- la surface de l'histogramme comprise entre la moyenne moins 2 écart-type et la moyenne plus 2 écart-type vaut 95% de la surface totale de l'histogramme.
- la surface de l'histogramme comprise entre la moyenne moins 3 écart-type et la moyenne plus 3 écart-type vaut 99 % de la surface totale de l'histogramme.

5. Interprétation des résultats

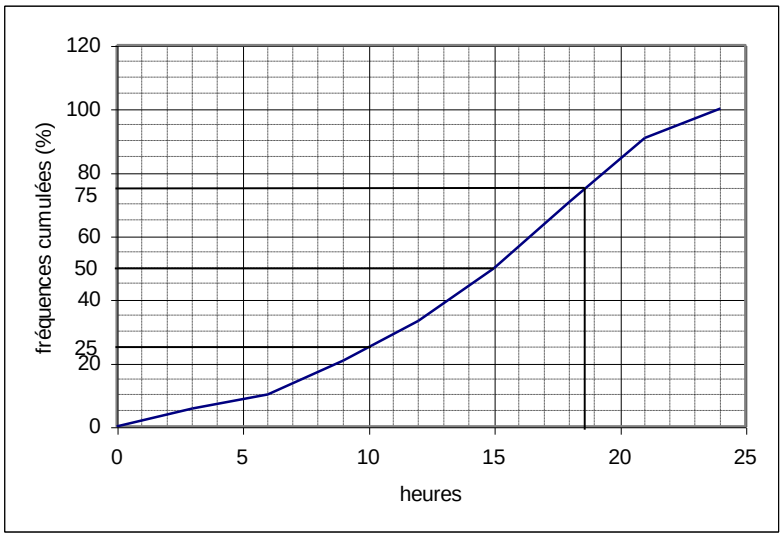
5.1 Tous les indicateurs n'ont pas le même intérêt

Selon la situation traitée les éléments à observer seront différents. Prenons un exemple : le tableau suivant décrit la répartition des accidents de la route selon les heures de la journée. On souhaite dégager les tendances essentielles des ces informations.

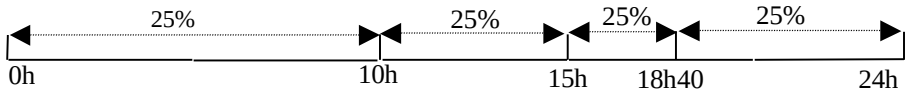
tranche horaire (en heures)	[0,3[	[3,6[	[6,9[	[9,12[	[12,15[	[15,18[	[18,21[	[21,24[
nombre d'accidents	8155	6258	15284	18006	23703	29759	29172	13022

On constate directement qu'un calcul de moyenne serait ici sans intérêt : affirmer que les accidents de la route ont lieu en moyenne à 14h04 n'a pas de sens. Cependant, les renseignements relatifs à la répartition sont plus intéressants.
Un graphique des fréquences cumulées nous permet de le voir :

tranche horaire	fréquences	fréquences cumulées
[0,3[	5.7	5.7
[3,6[	4.4	10.1
[6,9[	10.7	20.8
[9,12[	12.5	33.3
[12,15[	16.5	49.8
[15,18[	20.8	70.6
[18,21[	20.3	90.9
[21,24[	9.1	100



Interprétation :
La classe horaire [15,18[est la plus dangereuse (20,8% des accidents) : c'est le mode de la série.
Quelques points significatifs sur le graphique : les points d'ordonnées 50, 25 et 75.
Leurs abscisses (obtenues par lecture sur le graphique) nous permettent d'affirmer que les accidents se répartissent selon le schéma :



Et nous voyons ainsi tout l'intérêt des quartiles (10h et 18h40) et de la médiane (15h)
On peut résumer cette courte étude par

"Accidents : la période noire : 15h – 18h40
Dans la journée, si un accident sur deux se produit entre 10h et 18h40, c'est entre 15h et 18h40 qu'a lieu le quart des accidents."

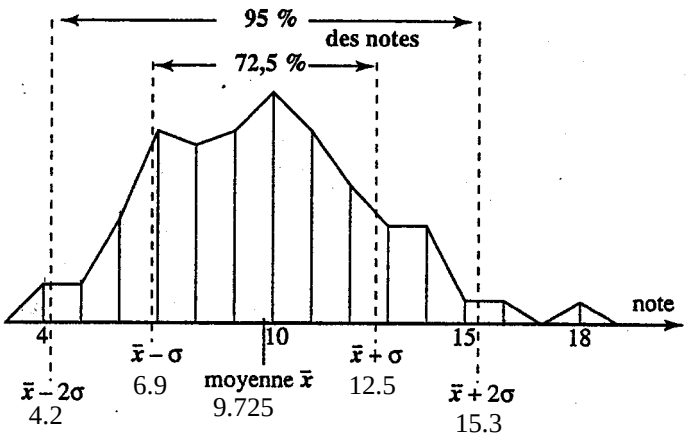
Remarquons cependant que dans d'autres cas, la recherche de la médiane n'a pas de sens.

5.2 Interprétation d'indicateurs statistiques

Lors d'un examen, les notes suivantes ont été obtenues (après remise en ordre) :

note	4	5	6	7	8	9	10	11	12	13	14	15	16	18
effectif	2	2	5	10	9	10	12	10	7	5	5	1	1	1

La calculatrice nous fournit les résultats suivants : $\bar{x} = 9.725$ et $\sigma = 2,75$
Un échantillon est considéré comme "normal" lorsque environ 30% des résultats sont hors de l'intervalle $[\bar{x}-\sigma, \bar{x}+\sigma]$ et 5% en dehors de l'intervalle $[\bar{x}-2\sigma, \bar{x}+2\sigma]$
Vérifions si l'échantillon obtenu est "normal".
$[\bar{x}-\sigma, \bar{x}+\sigma]$ correspond à l'intervalle [6.9…;12.5]
22 notes sont situées hors de cet intervalle, soit 27,5% de l'effectif des notes
$[\bar{x}-2\sigma, \bar{x}+2\sigma]$ correspond à l'intervalle [4.2;15.3]
4 notes sont situées hors de cet intervalle , soit 5% de l'effectif.
Nous pouvons schématiser cette situation sur le polygone des effectifs
L'écart-type apparaît ainsi comme une sorte d'étalon de dispersion



La "normalité" définie dans ce problème sera explicitée ultérieurement. Les statisticiens ont reconnu et répertorié des répartitions dites normales comme celles concernant les tailles d'individus, leur poids, les caractéristiques d'objets fabriqués par une machine…

6. Exercices généraux

6.1 Statistique à caractère qualitatif

Dans les deux situations qui suivent, calculer l'effectif total et les fréquences des différentes modalités. Tracer le diagramme à rectangles ainsi que le diagramme à secteurs.

6.1.1 Exercice 1

Dans la population des militaires en service actif à l'armée belge à la date du 15/07/83, on a relevé le nombre de ceux-ci dans chaque domaine.

Modalités	Effectifs
terre	32480
air	18760
mer	4760

6.1.2 Exercice 2

Dans un institut d'enseignement, on a relevé le nombre d'élèves dans chaque cycle:

Cycles	Effectifs
Observation	240
orientation	186
détermination	132

6.2 Statistiques à caractère discret.

Dans les exercices suivants:

1. Former : a) le tableau ordonné (si nécessaire)
   b) le tableau des effectifs (et/ou des fréquences)
   c) le tableau des effectifs cumulés (et/ou des fréquences cumulées)
2. Représenter : a) le diagramme en bâtons des effectifs (et/ou des fréquences)
   b) le polygone des effectifs cumulés (et/ou des fréquences cumulées)
3. Déterminer : a) les valeurs extrêmes, le(s) modes(s), la médiane, les quartiles, la moyenne.
   b) l'étendue, la variance, l'écart - type.
   c) préciser le sens de ces différents paramètres en fonction du contexte.

6.2.1 Exercice 1

Dans un carré de haricots on récolte 140 gousses et on compte le nombre de grains dans chacune des gousses cueillies. Les résultats obtenus sont :

5	9	3	6	5	4	5	1	6	2	7	6	5	3
4	7	6	5	8	6	4	5	6	4	7	7	6	9
5	4	5	6	5	6	3	5	7	2	7	6	7	9
8	6	3	7	5	4	6	3	5	4	6	2	6	4
10	6	3	6	6	2	6	4	7	5	6	5	6	7
1	9	5	6	7	5	6	5	7	5	5	8	6	7
9	6	7	8	6	2	4	7	6	4	3	6	8	6
7	4	6	3	5	6	7	5	4	9	6	5	5	4
6	7	5	6	5	7	10	6	7	5	6	8	4	6
5	2	5	4	6	5	4	5	6	5	4	8	5	3

6.2.2 Exercice 2

A la sortie d'une chaîne de fabrication, on tire chaque jour un lot de 1000 pièces prises au hasard et l'on poursuit cette opération pendant 100 jours. Dans chaque lot, on compte les pièces défectueuses. Au bout de 100 jours, on a obtenu:

Nbre de pièces defectueuses	0	1	2	3	4	5	6	7
Nbre de lots	5	16	23	21	17	9	6	3

6.2.3 Exercice 3

On a calculé le nombre de milliers de kilomètres parcourus par dix pneus de chacune des marques A et B avant usure. Les résultats suivants ont été obtenus :

A	25	28	26	34	30	24	28	22	27	23
B	31	29	24	26	21	32	27	29	26	24

6.3 Statistiques à caractère groupé.

Dans les exercices suivants

1. Former
   a) le tableau ordonné (si nécessaire)
   b) le tableau groupé des effectifs (et/ou des fréquences)
   c) le tableau groupé des effectifs cumulés (et/ou des fréquences cumulées)
2. Représenter
   a) l'histogramme des effectifs (et/ou des fréquences)
   b) le diagramme des effectifs cumulés (et/ou des fréquences cumulées)
3. Déterminer
   a) les valeurs extrêmes, la (les) classes(s) modale(s), la médiane, les quartiles, la moyenne
   b) l'étendue, la variance, l'écart – type
   c) préciser le sens de ces différents paramètres en fonction du contexte.

6.3.1 Exercice 1

La série suivante représente le quotient intellectuel de 100 enfants:

75	112	100	116	99	111	85	82	108	85
94	91	118	103	102	133	98	106	92	102
115	109	100	57	108	77	94	121	100	107
104	67	111	88	87	97	102	98	101	88
90	93	95	107	80	106	120	91	101	103
109	100	127	112	107	98	83	98	89	106
79	117	85	94	119	93	100	90	102	87
95	109	142	93	94	72	98	105	122	104
104	79	102	104	107	97	100	109	103	107
106	96	83	107	102	110	102	76	98	88

Grouper ce tableau en 9 classes, en prenant pour limites de classes 55, 65, 75,...

6.3.2 Exercice 2

Les élèves de deux classes différentes ont passé des tests cotés sur 100. Voici les résultats:

Classe A

56	42	46	48	50	18	46	52	40	
66	34	72	40	48	70	42	38	50	
38	52	52	46	50	70	46			

Classe B

27	47	37	64	17	42	50	30	66
43	38	22	62	48	52	28	52	57
83	77	43	72	30	83	40		

Grouper ces tableaux en prenant pour limites de classes 15, 25, 35 ...
Répondre ensuite aux questions et comparer les résultats.

6.3.3 Exercice 3

Pour comparer deux lots de blé cultivés, l'un avec engrais, l'autre sans engrais, on a mesuré les longueurs de quelques épis prélevés au hasard. Les résultats exprimés en centimètres et au millimètre près, sont rassemblés ci-dessous:

10.2	10.7	10.1	8.8	11.5
9.7	11.1	9.6	11.1	10.2
10.6	10.4	9.1	9.6	9.5
11.6	10.3	9.4	11.2	12.5
11.2	10.7	9.6	10.3	11.1
11.6	10.2	10.5	12.1	11.3
11.2	10.8	10.7	11	10.8
11.9	11.8	12.2	10.5	10.7
11.4	11.3	10.7	10.6	10.5

9.7	9.1	8.7	9.2	8.8
7.4	8.2	10.1	8.5	8.4
9.1	9.3	8.8	7.7	9.1
8.5	10.4	8	8.6	9.6
9.2	9.6	9	8.9	9.2
8.6	8.7	7.8	9.7	8.4
9.2	10.6	9.5	9.2	8.6
9	8.3	10.3	8.2	9.8
9	9.7	9.3	9.4	9.1

Grouper ces tableaux en prenant pour limites de classes 7 ; 7.5 ; 8 ; ...

6.3.4 Exercice 4

Les mesures de la taille des individus d'une population d'un petit batracien ont fourni les résultats suivants:

Taille en mm	Nombre
11 à 12	6
12 à 13	13
13 à 14	14
14 à 15	11
15 à 16	5
16 à 17	1
17 à 18	0

Taille en mm	Nombre
18 à 19	0
19 à 20	1
20 à 21	3
21 à 22	6
22 à 23	5
23 à 24	4
24 à 25	1

6.3.5 Exercice 5

Le tableau ci-contre nous donne le nombre de demandes d'admission, ventilé par âges, adressées par les tribunaux de la jeunesse à une institution de la région de Mons durant une période déterminée.

âges	nombre de demandes
[1,3[	1
[3,5[	6
[5,7[	3
[7,9[	2
[9,11[	4
[11,13[	6
[13,15[	19
[15,17[	25
[17,19[	26
[19,20]	10

6.3.6 Exercice 6

Voici le relevé des âges des habitants d'une commune à une date précise.

Ages (en années)	Nombre d'hab.
[0,10[	48
[10,20[	42
[20,30[	60
[30,40[	106
[40,50[	94
[50,60[	67
[60,70[	43
[70,100]	40

6.3.7 Exercice 7

Voici, relevé en 1981, un sondage effectué sur un échantillon de voitures de tourisme pour lesquelles on a observé le kilométrage parcouru au moment de la mise hors circulation

Kilométrage (en 10^3 km)	Nombre de voitures
[0,20[	400
[20,40[	650
[40,60[	850
[60,80[	800
[80,100[	1600
[100,120[	400
[120,140[	200
[140,200]	100

6.3.8 Exercice 8

Les valeurs suivantes donnent le poids en kg des 40 membres d'un club sportif.
Grouper ces données en classes d'extrémités 42, 50, 58, 66, 74, 80 avant d'en faire l'étude statistique complète.

70	66	55	69	52	60	42	61	59	48
65	50	69	56	62	70	61	58	61	76
51	63	60	66	57	62	80	68	52	49
64	52	72	46	67	57	68	72	54	77

6.4 Exercices variés.

6.4.1 Exercice 1

1) Interrogé sur les performances de sa voiture, du point de vue de la consommation, Mr Dupont a calculé à plusieurs reprises la quantité d'essence (en litres) consommée en 100 km. Il fournit les données suivantes,

présentées par ordre croissant. : 4.05 ; 4.12 ; 4.20 ;4.25 ; 4.63 ; 4.76 ; 4.81 ; 5.03 ; 5.35 ; 5.50 ; 5.63 ; 5.78 ; 5.91 ; 6.00 ; 6.00 ; 6.22 ; 6.34 ; 6.42 ; 6.65 ; 7.17 ; 7.17 ; 7.44 ; 7.44 ; 7.51 .

a) Regroupez ces valeurs en classes d'amplitude égale, les limites de la première classe étant [4.00 ; 5.00[.

b) Calculez deux mesures de dispersion.

2) Mr. Dupont dit de sa voiture qu'elle est "sobre et régulière". Justifiez cette appréciation en vous référant aux calculs que vous venez d'effectuer et dites à quelles mesures se rapportent les qualificatifs employés par Mr Dupont.

3) En supposant que Mr Dupont a payé son essence au prix constant de 32 fr/litre, combien paye-t-il, en moyenne, pour l'essence que sa voiture consomme en 100 km ?

4) Mr Durand s'est livré aux mêmes calculs que Mr Dupont. Sa voiture consomme en moyenne 8 litres aux 100 km, et l'écart-type de cette consommation est de .2.9 litres. Qualifiez en deux mots la voiture de Mr Durand, du point de vue de sa consommation.

6.4.2 Exercice 2

Une enquête est menée pour un magazine auprès des lycéens :

"Combien fumez-vous de cigarettes par jour ?

Le tableau ci-contre en donne les résultats :

Cigarettes :	Fréquences (Nb d'élèves)
0 à 5	35%
6 à 10	30%
11 à 15	10%
16 à 20	9%
21 à 25	2%
sans réponse	14%

a) Quelles raisons peuvent expliquer le pourcentage de "sans réponse" ?

Par la suite, on décide de ne pas tenir compte de cette donnée (autrement dit, on considère la nouvelle série statistique, déduite de la précédente en ne retenant que les lycéens ayant effectivement répondu)

b) Donnez le tableau des fréquences de cette nouvelle série et représentez son histogramme.

c) Quelle est la moyenne de la série obtenue?

6.4.3 Exercice 2

L'aptitude à la lecture a été étudiée chez les garçons de 7 à 11 ans. Le temps (secondes) mis pour lire un texte à haute et intelligible voix a été mesuré. Les résultats sont représentés dans le tableau ci-contre:

Durée de lecture	Nombre d'enfants
[25,30[	12
[30,35[	58
[35,40[	5
[40,45[	41
[45,50[	19
[50,55[	52
[55,60[	23
[60,65[	37
[65,70[	7
[70,75]	11

a)Quelle est la médiane de cette distribution ?

b)Quels en sont les quartiles ?

c) Quelle est la moyenne ?

d) Quel est l'écart - type ?

e) Que représente concrètement chacun de ces nombres ?

f) Ce groupe d'enfants vous paraît-il homogène en ce qui concerne la variable étudiée ? Justifiez votre réponse.

6.4.4 Exercice 4

Deux tireurs X et Y s'affrontent en vue d'une sélection lors d'une épreuve comportant vingt tirs sur cible :

Les résultats obtenus par les tireurs sont donnés par le tableau ci-contre.:

	50	30	20	10	0
X	4	6	5	4	1
Y	6	3	5	3	3

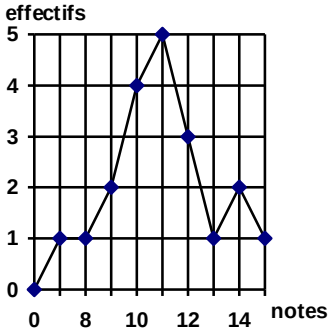
a) La moyenne par tir permet-elle de départager les deux concurrents ?

b) Reprendre la question précédente en ne tenant compte que des 10 meilleurs tirs.

c) Calculer l'écart - type de chacune des séries du tableau. Quel est le tireur le plus régulier ?

6.4.5 Exercice 5

Considérer le diagramme ci-contre indiquant les notes obtenues dans une classe à un devoir de statistique :
les affirmations suivantes sont vraies ou fausses ?
a) La classe est surchargée.
b) 40% des élèves ont une note inférieure ou égale à 10
c) La moyenne de la classe est égale à 11
d) Il y a autant de notes supérieures à la moyenne de la classe que de notes qui lui sont inférieures.



6.4.6 Exercice 6

Une machine remplit automatiquement des paquets de tabac. On prélève un échantillon de la production; après pesée, on obtient la distribution des masses des paquets suivante :

1. Calculer la moyenne et l'écart-type de la distribution des masses des paquets de tabac.
2. faire un nouveau tableau donnant les effectifs par classe d'amplitude 2 g et reprendre les calculs précédents avec la série ainsi obtenue.
3. Expliquer la différence (sensible) entre les écarts types trouvés aux questions 1 et 2
4. Cette machine est-elle fiable ?

Masse en grammes	effectifs
moins de 38	0
moins de 39	3
moins de 39.5	8
moins de 40	18
moins de 40.5	31
moins de 41	51
moins de 41.5	69
moins de 42	84
moins de 42.5	95
moins de 43	99
moins de 44	100
plus de 44	0

6.4.7 Exercice 7

On donne la répartition d'un groupe d'enfants par tailles (en cm)

1. Tracer l'histogramme de cette répartition.
2. Calculer la moyenne $\bar{x}$.
3. Calculer l'écart-type σ, puis le pourcentage d'enfants ayant une taille comprise entre : a) $\bar{x} - \sigma$ et $\bar{x} + \sigma$
   b) $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$
   Que peut-on en conclure ?

Taille (en cm)	Effectif
80 à moins de 90	3
90 à moins de 95	15
95 à moins de 100	22
100 à moins de 105	18
105 à moins de 110	12
110 à moins de 120	5

6.4.8 exercice 8

Une conserverie alimentaire fabrique des plats cuisinés mis en barquettes automatiquement. "les poids nets" de produit de 60 barquettes sont consignés ci-dessous (en grammes)

826	832	833	838	812	809	817	837	832	829	835	835	833	826	834	837	829	821	823	845
840	830	816	818	828	821	832	838	836	812	826	824	840	834	819	814	828	838	835	830
832	828	837	825	831	825	838	832	830	831	821	817	819	820	825	835	839	832	820	817

L'appareil de remplissage est en bon état de marche si $820\text{ g} \leq \bar{x} \leq 840\text{ g}$ et $\sigma \leq 10\text{g}$ et si la proportion de barquettes hors de l'intervalle $[\bar{x} - \sigma, \bar{x} + \sigma]$ ne dépasse pas 30%. Montrer que l'appareil est défectueux.

6.4.9 exercice 9

Dans un centre thermal, on a relevé la masse perdue (exprimée en kg) par les clients sur une durée de 15 jours.
Les valeurs obtenues sont reprises dans le tableau ci-contre

Masse (en kg)	Effectifs
[0, 2[	
[2, 4[	
[4, 6[	
[6, 8[	
[8, 12[	

a) Combien de personnes ont fréquenté ce centre durant ces 15 jours ?
b) Représente l'histogramme des effectifs
c) Représente le polygone des fréquences cumulées
d) Détermine la moyenne arithmétique, la classe modale et la médiane (de manière précise)
e) Détermine l'étendue, la variance et l'écart-type.
f) Si ce centre veut récompenser les 18% de ses clients qui ont perdu le plus de poids, détermine la masse minimale à perdre.

7. Statistique à 2 variables

Nous sommes souvent confrontés à des données entre lesquelles nous essayons d'établir des liens telles que :

- La taille et le poids d'un groupe d'individus.
- le budget vacances et les revenus des familles
- Le poids des récoltes et la durée d'ensoleillement ou la quantité de pluie reçue
- ……

Mais comment à partir de ces données, tirer des conclusions, exprimer le lien qui les unit ?
Prenons deux exemples qui nous serviront référence.

7.1 Exemple 1

Des élèves ont présenté un examen de mathématique et un examen de physique.
Les résultats sont les suivants :

Elèves N° :	1	2	3	4	5	6	7	8	9	10	11
Cote en math. (sur 20) : x_i	4	5	5	7	10	11	13	15	15	17	17
Cote en physique (sur 20): y_i	5	9	14	7	9	11	12	12	14	5	17

Au vu de ces résultats, on peut se poser différentes questions :
Les élèves forts en physique sont-ils forts en math ?
Y a-t-il un lien entre la cote de mathématique et celle de physique ? Et si oui, comment exprimer ce lien ?

7.2 Exemple 2

Un professeur de mathématiques a filmé avec une caméra numérique son fils en train de lancer un ballon. En regardant cet enregistrement avec arrêts sur images, il repère les données présentées ci-dessous.

durée (en s)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Hauteur (en cm)	84	121	149	167	175	174	163	143	114	75

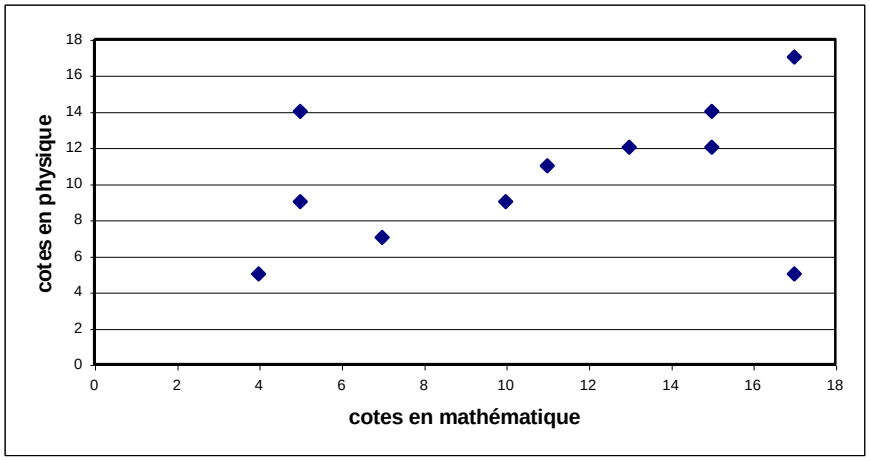
Comment, à partir de ces données déterminer une fonction h(t) qui exprime la hauteur du ballon en fonction du temps ?

8. Représentations graphiques

Pour se faire une meilleure idée des problèmes, représentons les données sur des graphiques.

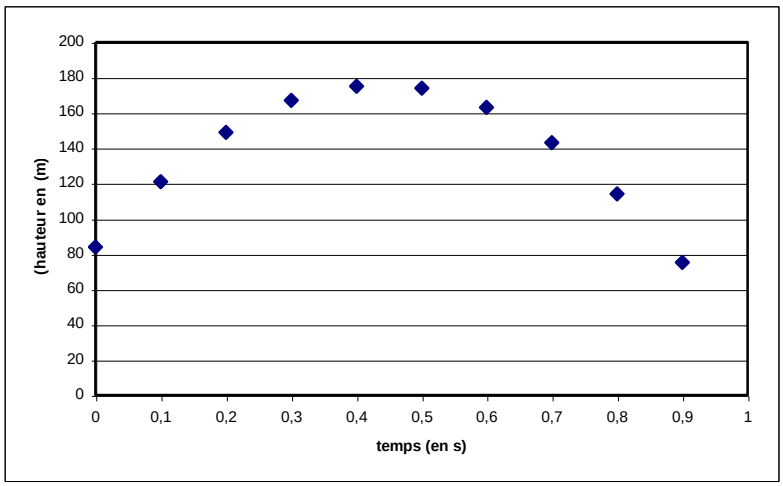
8.1 Exemple 1

Pour le premier exemple, nous obtenons le graphique suivant (nuage de points)



8.2 Exemple 2

En procédant comme pour le premier exemple, nous obtenons :



8.3 Observations :

En regardant ces graphiques, nous constatons que le type de fonction à ajuster sera différent selon les données : pour le premier exemple, si on peut ajuster une fonction, il s'agirait plutôt d'une fonction du premier degré tandis que dans le second, on choisirait une fonction du second degré.

8.4 Généralisation

Lorsque nous avons un nuage de points (x_i, y_i), différentes situations peuvent se présenter.

a) Les points sont disposés de façon quelconque : on dira que les caractères x et y sont indépendants.

b) Les points sont disposés autour d'une certaine courbe : on pourra faire un "ajustement graphique", c'est à dire tracer au mieux cette courbe.
La courbe la plus simple que l'on puisse obtenir est une droite. Parfois il s'agira d'une parabole, d'une fonction du troisième degré, ou d'autres fonctions que nous serons amenés à étudier ultérieurement.

c) Tracer cette courbe à main levée est très arbitraire, c'est pourquoi, nous allons développer des procédures de calcul.

9. Ajustement linéaire

Dans un premier temps, nous allons considérer des situations du genre du premier exemple, où le type de fonction à ajuster est une fonction du premier degré.

9.1.1 Ajustement graphique "à vue"

On trace une droite qui nous semble la plus près possible des points du nuage. En prenant deux points de la droite, nous pouvons obtenir rapidement son équation. Si nous utilisons cette technique en classe dans le premier exemple, nous constatons que plusieurs élèves peuvent obtenir des droites et donc des équations différentes. C'est donc une méthode rapide mais très approximative.

9.1.2 Droite de Mayer

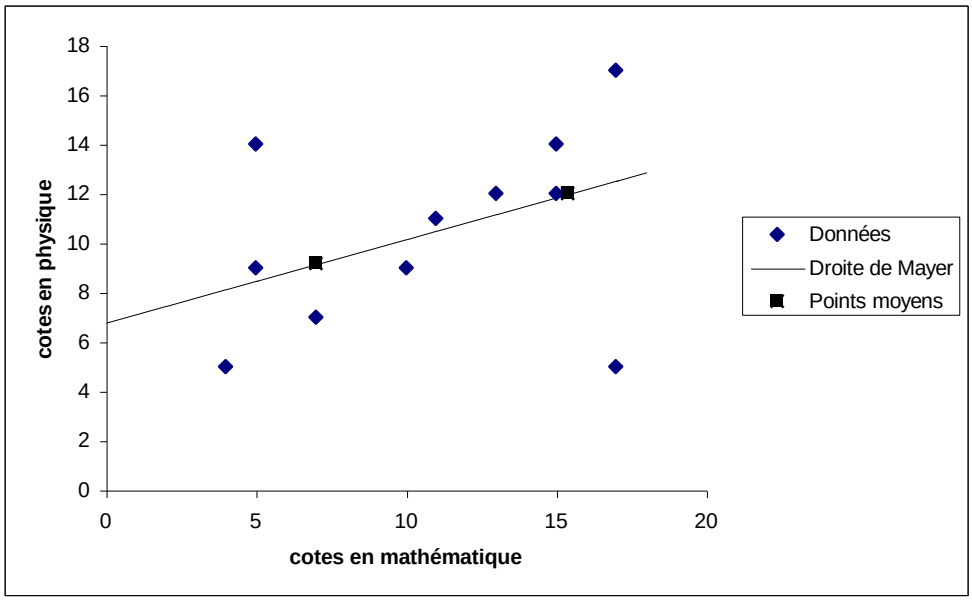
On ordonne les points de la série de manière croissante selon les xi et on divise le nuage en deux parties contenant le même nombre d'éléments (à une unité près si les données sont en nombre impair). Dans chaque sous-nuage, on calcule le point moyen. La droite qui passe par ces deux points est appelée droite de Mayer

Dans notre exemple, nous obtenons

$P_1\ (\bar{x}_1, \bar{y}_1) = \left(7, \frac{55}{6}\right)$: point moyen des 6 premiers points

$P_2\ (\bar{x}_2, \bar{y}_2) = \left(\frac{77}{5}, 12\right)$: point moyen des 5 derniers points

La droite P_1P_2 (droite de Mayer) a pour équation y = 0.3373x + 6.8056
Le graphique ci-dessous reprend à la fois les données et la droite de Mayer

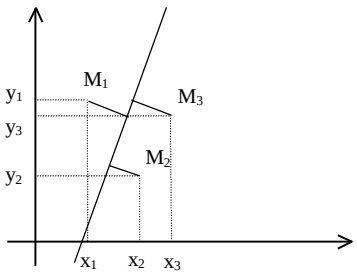


Ces 2 méthodes, si elles nous permettent de déterminer une droite qui s'approche des données ne nous permettent pas de déterminer "la meilleure droite d'approximation". Nous allons maintenant définir des critères qui vont permettre de décider quelle est cette "meilleure droite".

9.1.3 Ajustement par la méthode des moindres carrés.

Prenons une situation où nous avons n points : M_1 , M_2 , ...M_n de coordonnées respectives : (x_1 , y_1) (x_2 , y_2)...(x_n, y_n)
et soit $d \equiv y = ax + b$: la droite cherchée.

(le graphique a été réalisé pour le cas de 3 points)

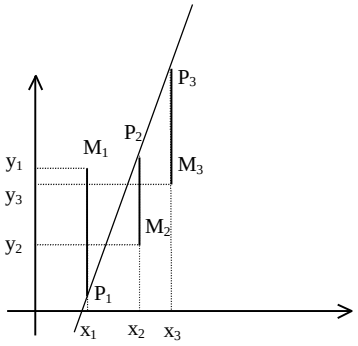


1ère idée : chercher a et b pour que la somme des distances des points M_i à la droite d soit la plus petite possible.

En réalisant ces calculs, on obtient une situation très compliquée. (Il faut alors minimiser $\sum_{i=1}^{n} \overline{M_i M'_i}$ où M'_i est la projection orthogonale de M_i sur la droite.

2ème idée : on peut aussi envisager de minimiser la somme des carrés des distances des M_i à la droite cherchée, mais là aussi cela reste très compliqué

3ème idée : On décompose le problème en 2 parties plus simples
Considérons d'abord P_i les points d'intersection des droites parallèles à OY menées par M_i
(le graphique ci-contre illustre une situation où on n'a que 3 points)
On veut que la somme des carrés des distances $\overline{M_iP_i}$ soit la plus petite possible: c'est pourquoi on appelle cette méthode la méthode des moindres carrés. La droite obtenue porte le nom de droite de régression de y en x.
Ensuite, on considérera les points d'intersection des droites parallèles à OX menées par M_i (voir graphique page suivante) et on minimisera la somme des carrés des distances $\overline{M_iQ_i}$ pour obtenir la droite de régression de x en y.



Si ces 2 droites sont proches, (confondues lorsque les points sont alignés), c'est que l'ajustement linéaire est "bon" pour la situation étudiée. Il faudra donc définir un coefficient qui mesure l'écart entre les deux droites : c'est le coefficient de corrélation linéaire.

** Détermination de la droite de régression de y en x : (pour information)

On veut minimiser la somme :$S = \overline{M_1P_1}^2 + \overline{M_2P_2}^2 + \ldots\ldots + \overline{M_nP_n}^2 = \sum_{i=1}^{n} \overline{M_iP_i}^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$

Nous allons minimiser cette somme dans le cas où on a trois points. (On peut généraliser cette démonstration.)

$S = \sum_{i=1}^{3} (y_i - ax_i - b)^2 = (y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + (y_3 - ax_3 - b)^2$

$= 3b^2 - 2b(y_1 + y_2 + y_3 - ax_1 - ax_2 - ax_3) + y_1^2 + y_2^2 + y_3^2 + a^2(x_1^2 + x_2^2 + x_3^2) - 2a(x_1y_1 + x_2y_2 + x_3y_3)$

En supposant a fixé, on cherche parmi toutes les valeurs de b celle qui minimise la somme. On voit qu'il s'agit d'un trinôme du second degré en b, le minimum est donc atteint pour

$b = \frac{2(y_1 + y_2 + y_3 - ax_1 - ax_2 - ax_3)}{3 \cdot 2} = \frac{y_1 + y_2 + y_3}{3} - a\frac{x_1 + x_2 + x_3}{3} = \bar{y} - a\bar{x}$

Nous avons ainsi exprimé b en fonction de a et de la moyenne des x_i et y_i
Et $d \equiv y = ax + \bar{y} - a\bar{x} \Leftrightarrow y - \bar{y} = a(x - \bar{x})$

Nous observons que le point moyen de coordonnée $(\bar{x}, \bar{y}) \in d$

Dans notre exemple : $\bar{x} = \frac{119}{11} = 10{,}81818$ et $\bar{y} = \frac{115}{11} = 10{,}4545$ et $b = 10{,}4545 - 10{,}8181\,a$

Si nous déterminons la valeur de a, nous aurons immédiatement celle de b.
En reportant la valeur de b trouvée dans la somme à minimiser, nous obtenons :
$S_a = [y_1 - ax_1 - (\bar{y} - a\bar{x})]^2 + [y_2 - ax_2 - (\bar{y} - a\bar{x})]^2 + [y_3 - ax_3 - (\bar{y} - a\bar{x})]^2$

$= [(y_1 - \bar{y}) - a(x_1 - \bar{x})]^2 + [(y_2 - \bar{y}) - a(x_2 - \bar{x})]^2 + [(y_3 - \bar{y}) - a(x_3 - \bar{x})]^2$

En ordonnant cette expression selon les puissances de a, nous obtenons :

$S_a = a^2[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2] - 2a[(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y})] + (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2$: une fonction du second degré en a. Elle admet un minimum pour

$$a = \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})}{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2}$$

$$= \frac{\frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})}{3}}{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2}{3}}$$

Le dénominateur de la fraction est la variance de la variable x et le numérateur est appelé covariance de x et y.

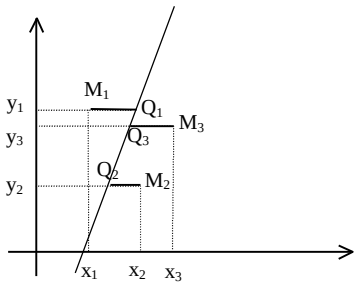
et donc $a = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Ce résultat peut se généraliser au cas de n observations.

Nous avons donc $a = \frac{\text{cov}(x,y)}{\text{var}(x)}$: coefficient angulaire de la droite de régression de y en x.

Cette droite minimise la $\sum_{i=1}^{n}\overline{M_iP_i}^2$. Comme nous avons vu qu'elle passe par le point moyen $(\bar{x}, \bar{y})$ elle a donc pour équation : $y - \bar{y} = a\,(x - \bar{x})$

De même, la droite de régression d' de x en y minimise la $\sum_{i=1}^{n}\overline{M_iQ_i}^2$



On démontre de même que son coefficient angulaire $a' = \frac{\text{var}(y)}{\text{cov}(x,y)}$ et que d' passe par le point moyen $(\bar{x}, \bar{y})$; elle a donc pour équation :
$y - \bar{y} = a'\,(x - \bar{x})$

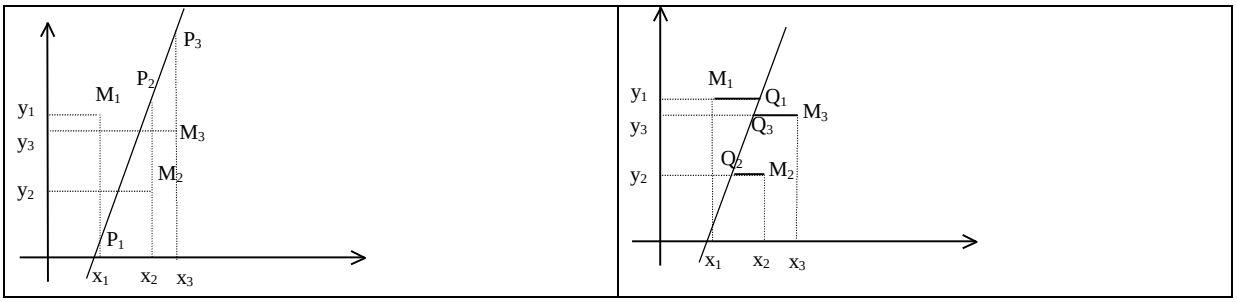
On définit $r^2 = \frac{a}{a'} = \frac{(\text{cov}(x,y))^2}{\text{var}(x)\ \text{var}(y)}$

D'où l'on déduit : le coefficient de corrélation linéaire $r = \frac{\text{cov}(x,y)}{\sigma(x)\,\sigma(y)}$

On peut montrer que $0 < |r| < 1$
Ce coefficient permet de mesurer "l'écart" entre d et d'. Si les 2 droites sont confondues, a = a' et r = 1.
Si la dépendance linéaire est mauvaise, |r| est éloigné de 1.
On considère que si $|r| < 0.87$, il n'y a pas de dépendance linéaire.
Il y a une corrélation positive entre les 2 variables lorsque les variations des deux variables se produisent dans le même sens (les pentes des droites de régression sont positives)
Il y a une corrélation négative entre les 2 variables lorsque les variations des deux variables se produisent dans le sens contraire (les pentes des droites de régression sont négatives)

Dans la pratique, ces calculs sont rarement effectués. La plupart des calculatrices scientifiques actuelles permettent de déterminer les droites de régression de façon très aisée.

9.2 Synthèse sur les droites de régression



1. Covariance

$$\text{cov}(x, y) = \frac{1}{n}\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

2. La droite de régression de y en x : $d \equiv y = ax + b$ minimise la $\sum_{i=1}^{n} \overline{M_i P_i}^2$

d ϶ le point moyen : $(\bar{x}, \bar{y})$ et $a = \dfrac{\text{cov}(x, y)}{\text{var}(x)} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \Rightarrow d \equiv y - \bar{y} = a\,(x - \bar{x})$

3. La droite de régression de x en y : $d' \equiv y = a'x + b'$ minimise la $\sum_{i=1}^{n} \overline{M_i Q_i}^2$

d' ϶ le point moyen : $(\bar{x}, \bar{y})$ et $a' = \dfrac{\text{var}(y)}{\text{cov}(x, y)} \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})} \Rightarrow d' \equiv y - \bar{y} = a'\,(x - \bar{x})$

4. Le coefficient de corrélation r

$r^2 = \dfrac{a}{a'} \Rightarrow r = \dfrac{\text{cov}(x, y)}{s(x)\,s(y)}$ $(0 < |r| < 1)$

Si $|r| < 0.87$: la dépendance linéaire est mauvaise.
Si $0.87 < |r| < 1$ l'ajustement linéaire est bon et une des droites de régression peut être prise comme ajustement.

9.3 Application

Revenons maintenant à l'exemple 1 proposé plus haut. Nous allons calculer les droites de régression pour cet exemple et le coefficient de corrélation linéaire.

x_i (cotes en math.)	y_i (cote en phys.)
4	5
5	9
5	14
7	7
10	9
11	11
13	12
15	12
15	14
17	5
17	17

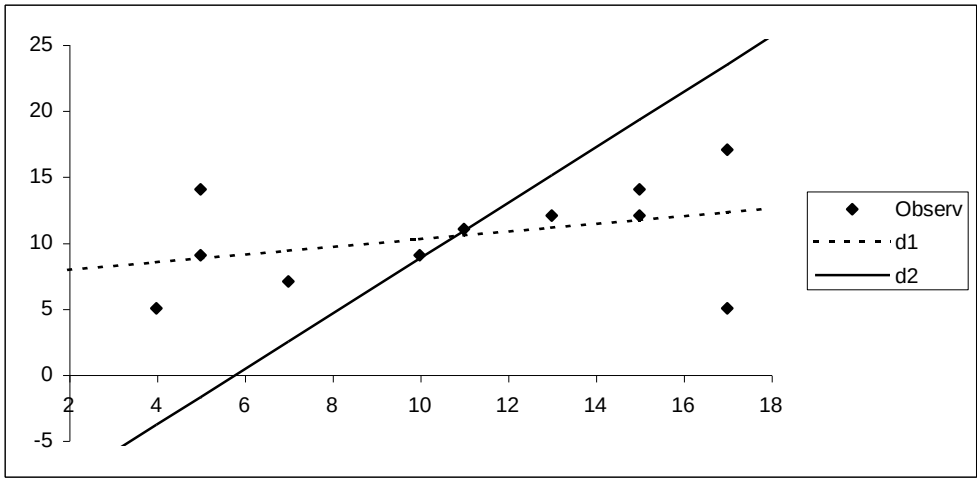
La calculatrice nous fournit directement les résultats :
$\bar{x} = 10.8181$ $\bar{y} = 10.4545$
ainsi que : $a = 0{,}28867$ $b = 7{,}3316$ et $r = 0.37098$

Nous avons ainsi la droite de régression de y en x :
$d_1 \equiv y = 0.28\,x + 7.33$
Pour déterminer la droite de régression de x en y, nous nous servirons de la relation :

$r^2 = \dfrac{a}{a'} \Leftrightarrow a' = \dfrac{a}{r^2}$ qui nous donne : $a' = \dfrac{0.28867}{0.37098^2} = 2.09749$

Or cette droite comprend le point moyen $(\bar{x}, \bar{y})$

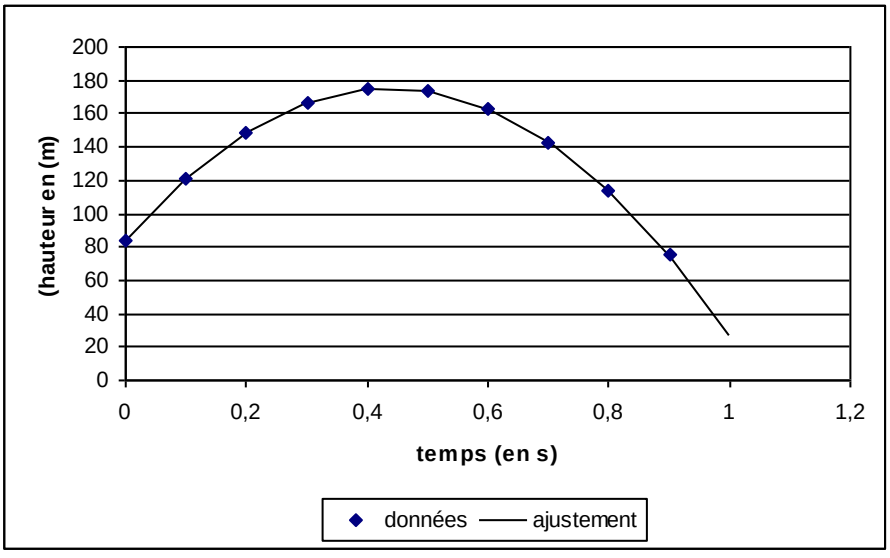
Et on a l'équation de $d_2 \equiv y - 10.45 = 2.1\,(x - 10{,}8) \Leftrightarrow y = 2.1\,x - 12.23$
Le graphique ci-dessous reprend les points (x_i, y_i) ainsi que les droites d_1 et d_2.



Le coefficient de corrélation r = 0.37 étant éloigné de 1, le décalage entre ces deux droites est important : on considère donc qu'il n'y a pas de dépendance linéaire.

10. Ajustement quadratique.

Dans le second exemple proposé, nous voulons ajuster une courbe du second degré.
Nous pouvons également procéder par tâtonnement : par exemple en résolvant un système de 3 équations à 3 inconnues (en prenant 3 des points des données) qui se ramène rapidement à un système de deux équations à 2 inconnues puisque l'on connaît l'ordonnée à l'origine. Comme dans le cas d'un ajustement linéaire, selon les points choisis, nous aurons plusieurs paraboles possibles.
Dans le cas qui nous occupe, on peut se référer aux formules de cinématique vues au cours de physique.
A nouveau, comme dans le point précédent, il va falloir choisir parmi toutes les possibilités la "meilleure courbe". Cependant, nous ne développerons pas ici les méthodes mathématiques qui permettent de choisir celle-ci, nous nous contenterons d'utiliser notre calculatrice.
Dans l'exemple proposé, nous obtenons : $h(t) = - 475\ t^2 + 417.015t + 84.218$ qui est comme nous pouvons le vérifier également assez proche des valeurs trouvées par tâtonnements.
Graphiquement, nous obtenons :



11. Exercices.

11.1 Exercice 1

Une firme productrice de véhicules spéciaux entreprend une étude statistique de ses coûts de production. Une collecte de données est résumée dans le tableau suivant :

Représenter graphiquement ces données déterminer les droites de régression associées à celles-ci.

Nombre d'unités produites : x_i	Coût global de production. : y_i (en 10^3 €.)
1000	350
2000	500
3000	575
4000	75
5000	925
6000	1025
7000	1175
8000	1275
9000	1350
10000	1500
11000	1575

11.2 Exercice 2

Le tableau ci-contre donne quelques chiffres sur le tourisme en Europe :

Pays	Nombre total de touristes arrivant (en millions.)	Recette totale (en 10^6 €)
Allemagne	4,9	450
Espagne	4,1	70
France	5,5	400
Italie	8,6	500
Suisse	4,6	250

On demande
1° de construire le nuage de points et de dire si un ajustement linéaire paraît vraisemblable.
2° d'établir les équations des droites de régression de y en x et de x en y
3° de dessiner ces droites
4° de calculer le coefficient de corrélation et d'indiquer ce que signifie ici ce coefficient.

11.3 Exercice 3

La direction commerciale d'une entreprise industrielle a augmenté régulièrement ses dépenses publicitaires pendant plusieurs années.
a) A partir du tableau ci-contre, comparer la progression du chiffre d'affaires avec les dépenses en déterminant les droites de régression et le coefficient de corrélation.
b) Tracer le nuage de points correspondant et les droites de régression.
c) Estimez les dépenses publicitaires à consentir pour atteindre 1 000 000€ comme chiffre d'affaire. Cette estimation est-elle fiable ?

Année	Dép. pub. (en €)	Chiffres d'affaires (en €)
1998	1830	881525
1999	1871	894275
2000	1998	919775
2001	1999	932 525
2002	2000	938 900
2003	2001	951 650
2004	2002	970 775

d) Peut-on estimer le chiffre d'affaires auquel on peut s'attendre si l'entreprise augmente son budget publicitaire jusque 3 000 € ? Justifier

11.4 Exercice 4

But : doser les nitrites dans l'eau afin de déterminer l'indice de pollution organique de l'eau avant et après lagunage (station de Sart-Bernard) (Expériences réalisées dans les classes de biologie appliquée du collège)
On établit donc une gamme étalon en nitrates allant de 0 à 25 mg de nitrates par litre.
Les résultats des mesures étalons sont donnés dans le graphique ci-contre.

Concentration (mg/l)	Absorbance
0	0
5	0.14
10	0.28
15	0.36
20	0.55
25	0.68

a) Donner une expression de l'absorbance en fonction de la concentration (droite de régression)
b) Exprimer la concentration en fonction de l'absorbance.
b) Si l'absorbance vaut 0.40, estimer la valeur de la concentration en nitrates.

11.5 Exercice 5

Dans un ancien rapport sur la conjoncture économique, on relève le tableau ci-contre :

a) Déterminer les droites de régression de cette série statistique.

b) Calculer son coefficient de corrélation ? Que pouvez-vous conclure ?

c) Si l'indice des prix vaut 155, peut-on prévoir le taux des salaires ? Cette prévision est-elle fiable ? Justifier.

France	taux des salaires horaires	indice d'ensemble des prix de détail
Mars 1950	105	105.3
Juin 1950	107	103.9
Sept. 1950	114	109.2
Déc. 1950	120	113.6
Mars 1951	127	119.7
Juin 1951	138	127
Sept. 1951	156	131.6
Déc. 1951	160	141.3
Mars 1952	162	146.9
Juin 1952	163	142.7
Sept. 1952	164	147.9

11.6 Exercice 6

Un test permet de mesurer l'aptitude à la lecture d'enfants en fonction de leur âge.
x_i représente l'âge et y_i la durée de lecture (en secondes)

a) Déterminer les équations des droites de régression

b) L'hypothèse : plus l'âge augmente, plus la durée diminue est-elle vérifiée ? Justifier.

x_i	y_i
7	58
8	59
9	46
10	33
11	31
11	33
9	44
7	74
7	60
8	60
10	41
10	31

11.7 Exercice 7

Une société a mis au point un produit. Une enquête menée auprès de 500 personnes a montré une relation entre le prix x proposé (en €) pour ce produit et le nombre de clients disposés à l'acheter à ce prix. Les résultats de cette enquêtes sont donnés dans le tableau ci-contre.

a) Construire le nuage de points

b) Déterminer les droites de régression et les tracer sur le graphique.

c) Interpréter le coefficient de corrélation.

d) Si on proposait le produit à 30€, à combien peut-on estimer le nombre de personnes qui en achèteraient ? Cette estimation est-elle fiable ? Pourquoi ?

Prix (en €)	Nombre de clients
40	60
35	80
32	130
28	200
24	240
20	350
16	390
12	420
10	440
8	500

11.8 Exercice 8

Prenons la situation fictive où l'on mesure la productivité d'un groupe de travailleurs de huit heures à quinze heures :
Dans ce tableau, on voit la productivité croître tout au long de la journée pour diminuer sous l'influence de la fatigue en fin de journée.
On aimerait savoir si la productivité est fonction de l'heure de la journée.
Représenter graphiquement la situation et utiliser l'ajustement le plus approprié.

Heure du jour	Productivité
8	12,0
9	13,0
10	14,0
11	14,4
12	14,8
13	14,5
14	14,0
15	13,0