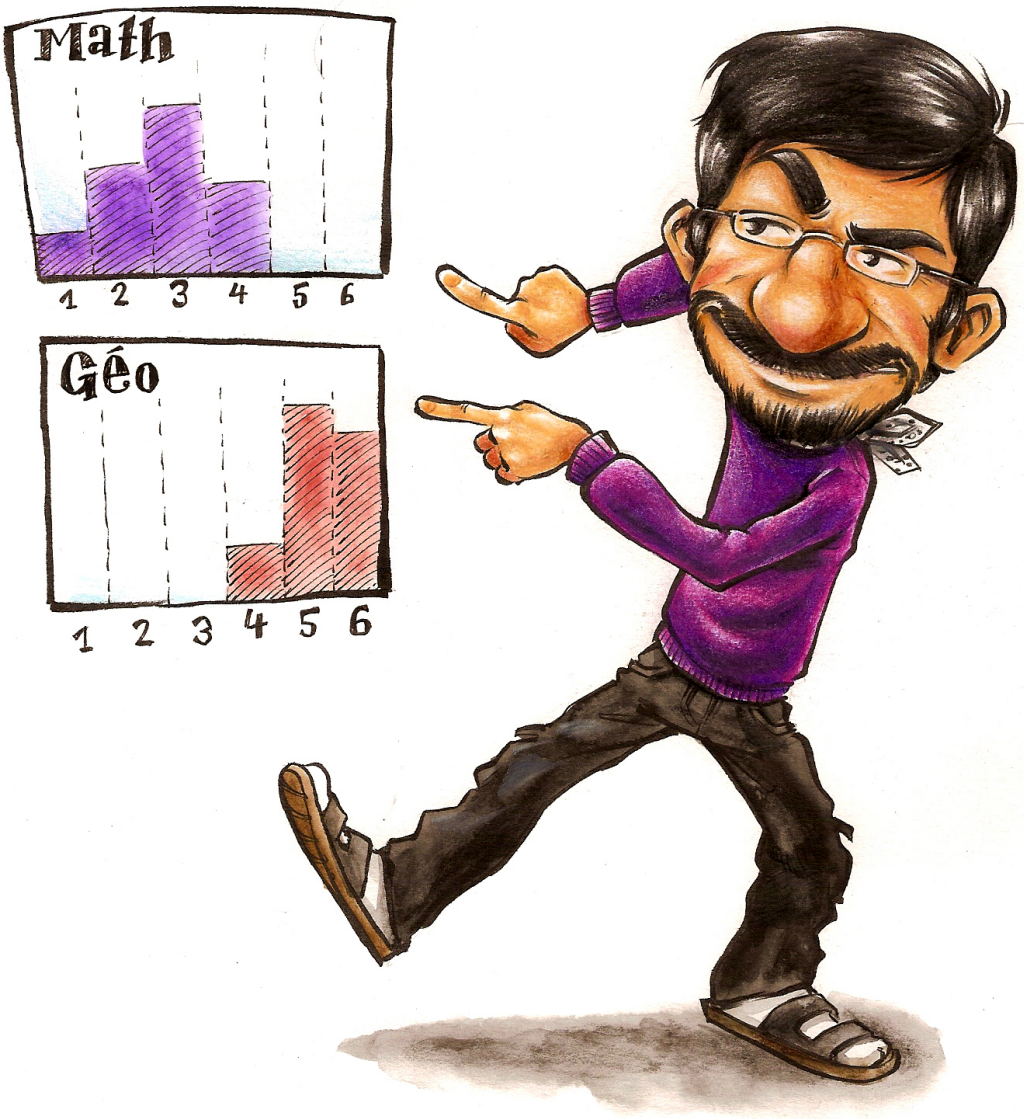
Statistiques



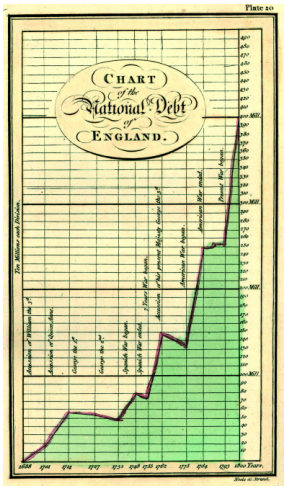
Didier Müller, août 2010

www.apprendre-en-ligne.net

Table des matières

1. Statistique descriptive

1.1. Un peu d'histoire



Comment interpréter « l'avalanche de chiffres » de la réalité sans outils théoriques ? L'humanité a mis fort longtemps avant de découvrir des procédés de calcul efficaces et des représentations pertinentes. Depuis, ces outils ont envahi tous les domaines de la connaissance. Il semble que les premiers paramètres de position qui aient été utilisés soient le mode, valeur apparaissant le plus fréquemment, et le « milieu de l'intervalle défini par les valeurs extrêmes ». La moyenne arithmétique apparaît clairement dans l'œuvre de l'astronome danois **Tycho Brahé** (1546-1601) qui, en constituant un ensemble de données sur le mouvement des planètes, permit à **Kepler** de formuler ses lois. En 1722, Roger **Cotes**, qui dispose d'observations qui ne sont pas toutes aussi fiables, propose d'utiliser une moyenne pondérée dont les coefficients sont inversement proportionnels à la dispersion des erreurs d'observations. On peut noter que la médiane voit naître son intérêt à la même époque, en 1757. La variance naît au 19e siècle avec les moindres carrés. **Gauss** lui préfère l'écart-type.
La représentation graphique quantitative trouve son origine dans la construction de cartes géographiques. Les plus anciennes datent d'environ 6000 ans, gravées sur des tablettes d'argile, en Mésopotamie. Les graphiques statistiques sont plus récents. William **Playfair** (1759-1823) publiera à Londres des ouvrages dans lesquels on trouve des graphiques de grande qualité (voir ci-contre) et entre autres le premier diagramme en barres connu ainsi que, une peu plus tard, le premier diagramme en secteurs.

1.2. Vocabulaire

Exemples de caractère d'une population :

- durées de vie d'ampoules
- poids de poulets d'élevage
- notes de math des élèves d'une classe

En statistique, on désigne par **population** tout ensemble d'objets de même nature. Ces objets présentent tous un certain **caractère** qu'il s'agit d'étudier pour en révéler les tendances principales. Lorsque la population est trop vaste pour l'étudier dans son ensemble, on en prélève au hasard un **échantillon** que l'on étudie. La taille de cet échantillon devra bien sûr être suffisamment grande pour pouvoir tirer des conclusions sur la population totale. Le caractère étudié est soit de nature **discrète** (il ne peut prendre que des valeurs réelles isolées, par exemple les notes entre 1 et 6 évaluées au demi-point), soit de nature **continue** (il peut prendre toute valeur d'un certain intervalle réel, comme la vitesse d'une voiture).
Les **tableaux** et les **graphiques** donnent une bonne idée de la manière dont un caractère est distribué, mais on cherche souvent à illustrer cette **distribution** de manière beaucoup plus sommaire par quelques nombres caractéristiques. Parmi ceux-ci, les **mesures de tendance centrale** (aussi appelées **paramètres de position**) jouent un rôle essentiel. La plus connue est la **moyenne**, mais on utilise aussi la **médiane** ou le **mode**.
Les mesures de tendance centrale ne suffisent pas à donner une idée de la manière dont les valeurs sont distribuées au voisinage de ces valeurs centrales. Aussi est-il utile d'introduire une **mesure de la dispersion**. La plus utilisée est **l'écart-type**. Dans le cas continu, **l'intervalle semi-interquartile** est aussi très fréquent.

1.3. Cas discret

On utilisera cet exemple pour illustrer les notions de ce paragraphe.

Dans une classe de 26 élèves, la maîtresse a relevé les notes suivantes :

4 4 5 3 1 5 4 6 2 4 3 5 5 5 0 4 5 6 3 3 5 2 5 4 4 3

Afin d'y voir plus clair, elle regroupe les notes dans un tableau. Dans la première colonne, elle numérote les 7 **observations** possibles, dans la deuxième, elle inscrit les **valeurs** de ces observations (les notes), et dans la dernière elle note les **effectifs**, i.e. le nombre de fois qu'apparaît chaque valeur.

Les premières statistiques sont probablement les recensements effectués à propos des individus et de leurs biens, il y a 4'500 ans en Mésopotamie et en Égypte.

De nos jours, les sondages d'opinion sont courants. Les statistiques sont très utilisées par les assurances.

Tableau 1

	Notes	*Élèves*
Observations i	Valeurs x_i	Effectifs n_i
1	0	1
2	1	1
3	2	2
4	3	5
5	4	7
6	5	8
7	6	2
		Effectif total : $n=\sum_{i=1}^{7} n_i=26$

Notation : $\sum_{i=1}^{N} x_i = x_1 + x_2 + \ldots + x_N$

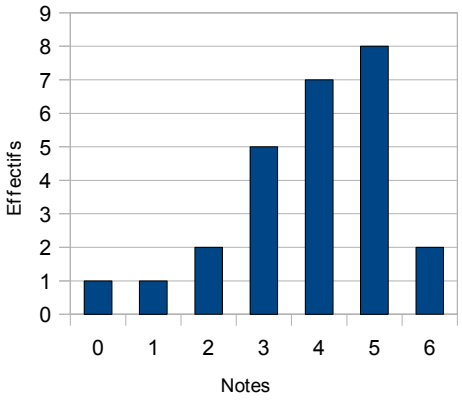
Exercice 1.1

Avec les données du tableau ci-dessus, calculez les expressions suivantes :

a. $\sum_{i=2}^{5} x_i$ **b.** $\sum_{k=1}^{6} n_k$ **c.** $\sum_{i=1}^{4} n_i x_i$ **d.** $\sum_{i=1}^{4} n_i \sum_{j=1}^{4} x_j$

Représentations graphiques

Les deux représentations graphiques les plus courantes sont l'**histogramme** (diagramme en bâtons) et le **diagramme à secteurs** (communément appelés « camemberts »).
Les deux graphiques suivants sont dessinés d'après les données présentées dans le tableau 1.



Histogramme

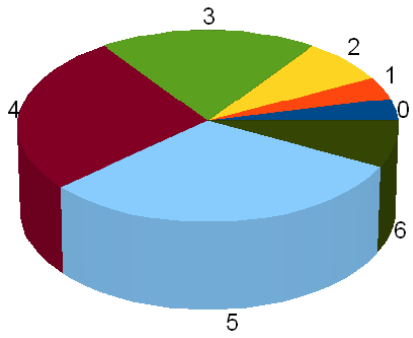
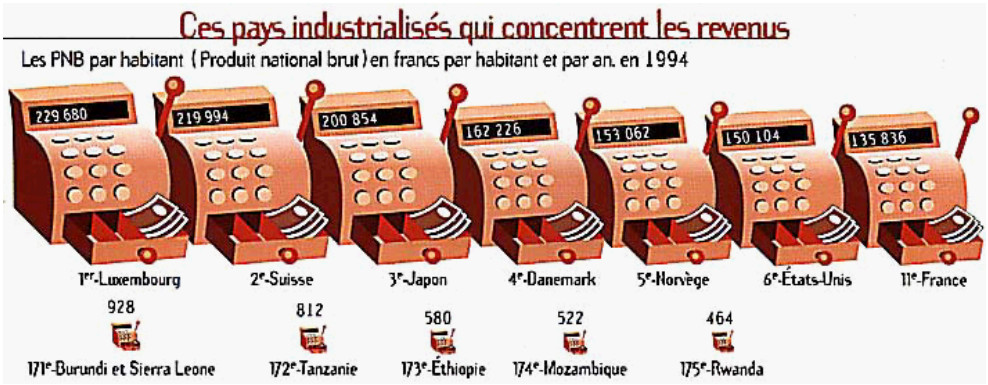


Diagramme à secteurs

On peut évidemment ajouter un côté « artistique » aux graphiques, comme dans l'exemple ci-dessous :



Moyenne

(mesure de tendance centrale)

La **moyenne** est la plus connue des mesures de tendance centrale. Elle s'obtient en divisant la somme des valeurs par le nombre de valeurs (n) :

$$\bar{x} = \frac{\sum_{i=1}^{7} n_i x_i}{n}$$

En utilisant les données du tableau 1, on trouve :

$$\bar{x} = \frac{1\cdot 0 + 1\cdot 1 + 2\cdot 2 + 5\cdot 3 + 7\cdot 4 + 8\cdot 5 + 2\cdot 6}{26} = \frac{100}{26} = 3.846$$

Remarque La moyenne est influencée par toutes les valeurs et est malheureusement très sensible aux valeurs extrêmes, au point d'en perdre parfois une bonne partie de sa représentativité, surtout dans des échantillons de petite taille. Ainsi la moyenne des six salaires mensuels suivants

3'500 4'200 4'600 5'000 6'200 36'500

est égale à 10'000 (!), alors qu'un seul salaire dépasse cette moyenne.

Variance et écart-type

(mesure de dispersion)

Si l'on désire se faire une idée de la manière dont les valeurs du caractère s'écartent de la moyenne $\bar{x}$ de ce caractère, on calcule la moyenne des écarts quadratiques :

$$\nu = \frac{\sum n_i (x_i - \bar{x})^2}{n} = \frac{\sum n_i x_i^2}{n} - \bar{x}^2$$

La deuxième expression est plus agréable pour les calculs.

ν est la **variance** de l'échantillon. L'**écart-type** σ est la racine carrée de la variance.

$$\sigma = \sqrt{\nu}$$

Vos calculatrices comprennent des touches spéciales pour calculer efficacement la moyenne et l'écart-type. Consultez votre mode d'emploi !

En utilisant les données du tableau 1, on trouve :

$\bar{x} = \frac{100}{26} = 3.846$; $\nu = \frac{438}{26} - 3.846^2 = 16.846 - 14.793 = 2.053$. D'où $\sigma = \sqrt{\nu} = 1.433$.

Remarque Quand on calcule la variance d'un échantillon (et non de la population entière), le dénominateur est n–1.

Exercice 1.2

Les trois élèves suivants ont 4 de moyenne. Et pourtant, ils sont très différents. Calculez l'écart-type de leurs quatre notes. Que constatez-vous ?

a. 4 4 4 4 **b.** 2 2 6 6 **c.** 2 3 5 6

Médiane

(mesure de tendance centrale)

On **trie** tout d'abord les n valeurs par ordre croissant :

0 1 2 2 3 3 3 3 3 4 4 4 **4 4** 4 4 5 5 5 5 5 5 5 5 6 6

La **médiane** est simplement la valeur qui se trouve au milieu : $\tilde{x} = x_{\frac{n+1}{2}}$.

Si n est pair, on prend la moyenne des deux valeurs du milieu : $\tilde{x} = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$.

Avec les données du tableau 1, $\tilde{x} = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) = \frac{1}{2}(x_{13} + x_{14}) = \frac{4+4}{2} = 4$.

Remarque La médiane n'est pas affectée par les valeurs extrêmes de la distribution.

Intervalle semi-interquartile

(mesure de dispersion)

Méthode de calcul

1. Trier les données dans l'ordre croissant.
2. Diviser les données en deux groupes de taille égale : le groupe *A* avant la médiane et le groupe *B* après la médiane (si l'échantillon de départ a une taille impaire, rajouter la médiane en tête du groupe *B*).
3. Calculer la médiane du groupe *A*, que l'on appellera Q_1.
4. Calculer la médiane du groupe *B*, que l'on appellera Q_3.
5. L'intervalle semi-interquartile (*isi*) vaut : $isi = \frac{Q_3 - Q_1}{2}$

Remarque :
par convention, $Q_2 = \tilde{x}$

Reprenons les données du tableau 1 :

Groupe *A*	Groupe *B*
0 1 2 2 3 3 **3** 3 3 4 4 4 4	4 4 4 5 5 5 **5** 5 5 5 5 6 6
$Q_1 = 3$	$Q_3 = 5$

$$isi = \frac{5-3}{2} = 1$$

Mode

(mesure de tendance centrale)

Le **mode** est par définition la valeur la plus fréquente dans une série de données.

En lisant le tableau 1, on constate que, dans cet exemple, le mode vaut 5.

Remarques

Le mode n'est pas affecté par les valeurs extrêmes de la distribution.
Selon la série de données, il peut y avoir plusieurs modes.

Exercice 1.3

Utilisez les touches spéciales de votre machine pour calculer la moyenne et l'écart-type.

Lors d'une journée, on a relevé les âges de 20 personnes venant se présenter à l'examen théorique du permis de conduire :

18	19	19	23	36	21	57	23	22	19
18	18	20	21	19	26	32	19	21	20

Calculez la moyenne, la médiane, le mode, la variance, l'écart-type et l'intervalle semi-interquartile de ces valeurs.

Exercice 1.4

Au laboratoire de physique, une série de mesures de l'accélération de la pesanteur terrestre a donné les résultats suivants :

9.95 9.85 10.13 9.69 9.47 9.98 9.87 9.46 10.00

Calculez la moyenne et l'écart-type des résultats.

Exercice 1.5

Le professeur de maths m'a dit : « C'est bien ; disons plutôt que c'est pas mal : tu as 4.5 de moyenne sur les cinq notes du semestre ». Sachant qu'aux quatre premières j'ai eu 5.2, 3.1, 4.4 et 4.2, quelle est ma note à la dernière épreuve ?

Exercice 1.6

41'250'000 personnes d'un pays ont atteint leur taille définitive (1.67 mètres en moyenne). Si l'on vous dit que, dans ce pays, la femme moyenne mesure 1.61 mètres et l'homme moyen 1.74 mètres, sauriez-vous en déduire de combien le nombre de femmes dépasse le nombre d'hommes dans ce pays ?

Exercice 1.7 (exercice de classe)

Chaque élève de la classe est prié de relever le prix de trente articles **différents** choisis **au hasard**, soit en se promenant dans un grand magasin, soit en parcourant un catalogue de vente par correspondance. Il notera ensuite combien de fois apparaît chaque premier chiffre significatif (le chiffre tout à gauche, 0 excepté), i.e. combien de fois le prix des articles commence par un 1, par un 2, …, et par un 9.

Jouez le jeu ! Les résultats seront rassemblés et analysés en classe.

1.4. Cas continu

Lorsqu'il y a **trop de valeurs discrètes**, ou lorsque le caractère de la population est de **nature continue**, on regroupe les valeurs en **classes** de même amplitude.

Lors d'une course de vitesse, les 40 participants ont mis les temps ci-contre pour effectuer le parcours.
On représente ces données par un histogramme dans lequel chaque classe (ici d'**amplitude** 2) se voit attribuer un rectangle dont l'aire est proportionnelle à l'effectif de la classe.

Tableau 2

Temps (*classes*)	Centres des classes x_i	Effectifs n_i
[43-45[	44	2
[45-47[	46	3
[47-49[	48	7
[49-51[	50	11
[51-53[	52	8
[53-55[	54	6
[55-57[	56	3
		$n = 40$

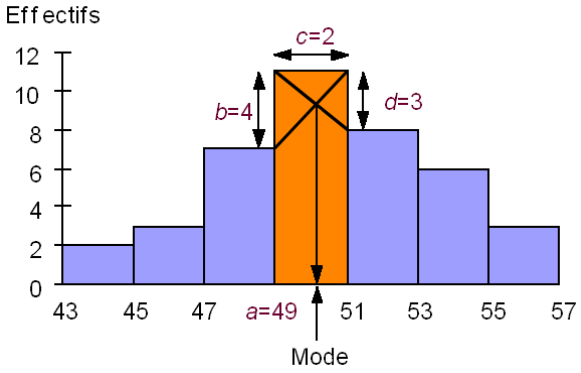
Mode

Dans le cas continu, le mode se trouve dans la classe ayant le plus grand effectif (la **classe modale**).

Il se calcule sur l'histogramme ainsi : **mode** = $a + c \cdot \frac{b}{b+d}$

Ci-dessous : **mode** = $49 + \frac{2 \cdot 4}{4+3} = 50.14...$

Il peut y avoir plusieurs classes modales, donc plusieurs modes.



Fréquences et fréquences cumulées

Il est souvent intéressant de faire figurer dans un tableau statistique, pour chaque valeur (ou pour chaque classe) x_i que peut prendre le caractère, la proportion f_i des individus qui présentent cette valeur x_i . Ces proportions sont appelées **fréquences**.

Si n est l'effectif total, alors par définition $f_i = \frac{n_i}{n}$.

La **fréquence cumulée** $F(x)$ est la proportion des individus qui présentent des valeurs x_i inférieures ou égales à x. Elle se calcule en additionnant toutes les fréquences f_i correspondant aux x_i tels que $x_i \leq x$.

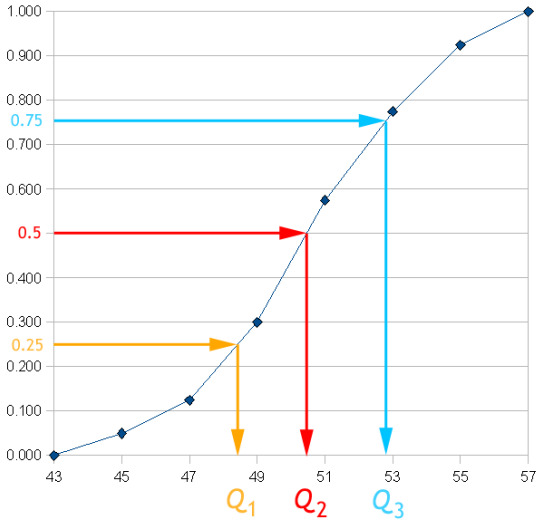
Ce tableau représente les vitesses de 40 voitures mesurées dans un village.

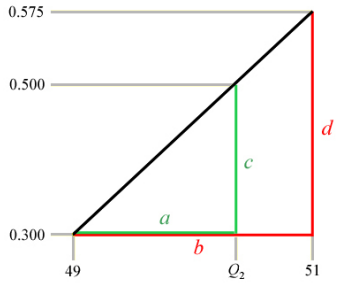
Tableau 3

Classes (temps)	Centres des classes x_i	Effectifs n_i	Fréquences f_i	Fréquences cumulées $F(x_i+1)$
[43-45[	44	2	2/40 = 0.050	2/40 = 0.050
[45-47[	46	3	3/40 = 0.075	5/40 = 0.125
[47-49[	48	7	7/40 = 0.175	12/40 = 0.300
[49-51[	50	11	11/40 = 0.275	23/40 = 0.575
[51-53[	52	8	8/40 = 0.200	31/40 = 0.775
[53-55[	54	6	6/40 = 0.150	37/40 = 0.925
[55-57[	56	3	3/40 = 0.075	40/40 = 1.000
		$\Sigma = 40$	$\Sigma = 1$	

On obtient le **polygone des fréquences cumulées** ci-dessous :

Le polygone des fréquences cumulées commence à une ordonnée de 0 et finit en 1.





Médiane

La médiane se calcule en utilisant le polygone des fréquences cumulées. Il faut repérer quel segment coupe la droite horizontale d'ordonnée 0.5, puis calculer la médiane par proportionnalité (grâce au théorème de *Thalès*).

$$\frac{a}{b}=\frac{c}{d} \Rightarrow \frac{Q_2-49}{51-49}=\frac{0.5-0.3}{0.575-0.3} \Rightarrow Q_2=49+2\cdot\frac{0.2}{0.275}=50.45\ldots$$

Intervalle semi-interquartile

F étant la fonction représentative du polygone des fréquences cumulées, on appelle respectivement premier, deuxième et troisième quartile les valeurs Q_1, Q_2 et Q_3 telles que

$$F(Q_1)=\frac{1}{4}\ ;\ F(Q_2)=\frac{2}{4}\ ;\ F(Q_3)=\frac{3}{4}$$

On voit que l'intervalle $[Q_1; Q_3]$ contient le 50% des valeurs de l'échantillon.

L'intervalle semi-interquartile est égal, par définition, à la moitié de la longueur de cet intervalle :

$$isi=\frac{Q_3-Q_1}{2}$$

Q_1 et Q_3 se calculent de manière similaire à la médiane.

$$\frac{Q_1-47}{49-47}=\frac{0.25-0.125}{0.3-0.125} \Rightarrow Q_1=47+2\cdot\frac{0.125}{0.175}\approx 48.428$$

$$\frac{Q_3-51}{53-51}=\frac{0.75-0.575}{0.775-0.575} \Rightarrow Q_3=51+2\cdot\frac{0.175}{0.2}=52.75$$

$$isi=\frac{52.75-48.428}{2}\approx 2.161$$

Moyenne et écart-type

Dans le cas continu, la moyenne et l'écart-type se calculent comme dans le cas discret en utilisant comme valeurs les centres de classes. **Ces mesures changeront légèrement selon la manière dont on aura formé les classes.**

Remarque Si on utilise la moyenne pour mesurer la tendance centrale, on lui associera l'écart-type pour mesurer la dispersion. Si par contre on utilise la médiane, on lui associera l'intervalle semi-interquartile.

Exercice 1.8

Lors d'un contrôle de police sur l'autoroute, un agent a relevé les vitesses suivantes (arrondies à l'entier inférieur ou égal) :

117	134	130	113	127	125	98	110	124	122	126	101
106	121	121	104	124	117	109	128	134	146	111	139
123	124	130	123	120	133	111	143	145	111	110	119
114	104	126	99	140	105	119	134	128	119	137	109
122	130	92	104	113	130	120	84	166	138	129	119

a. Groupez ces données par classes : [80-90[, [90-100[, etc.
b. Dessinez le diagramme à secteurs correspondant.
c. Calculez le mode, la médiane et l'intervalle semi-interquartile.

Exercice 1.9

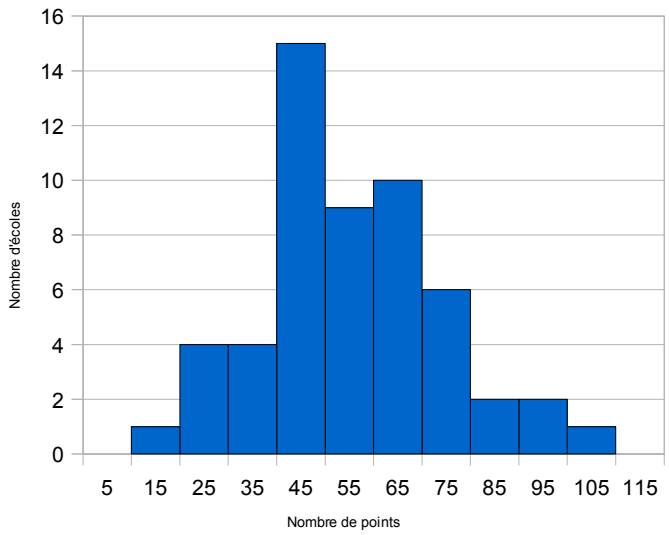
Les salaires mensuels payés aux ouvriers d'une entreprise se répartissent comme suit :

4	ouvriers gagnent entre 2400 et 2700 francs
21	ouvriers gagnent entre 2700 et 3000 francs
104	ouvriers gagnent entre 3000 et 3300 francs
163	ouvriers gagnent entre 3300 et 3600 francs
121	ouvriers gagnent entre 3600 et 3900 francs
57	ouvriers gagnent entre 3900 et 4200 francs
22	ouvriers gagnent entre 4200 et 4500 francs
10	ouvriers gagnent entre 4500 et 4800 francs

a. Faites un tableau en vous inspirant du tableau 3.
b. Dessinez l'histogramme et le polygone des fréquences cumulées.
c. Calculez le mode, la médiane et l'intervalle semi-interquartile.
d. Calculez le salaire mensuel moyen et l'écart-type.

Exercice 1.10

Au concours de *Mathématiques sans Frontières*, le nombre de points obtenus par les écoles de Suisse se répartit selon l'histogramme suivant :



a. Calculez la moyenne de cette série.
b. En utilisant l'histogramme, trouvez le pourcentage des écoles qui ont moins de 64 points.

Exercice 1.11

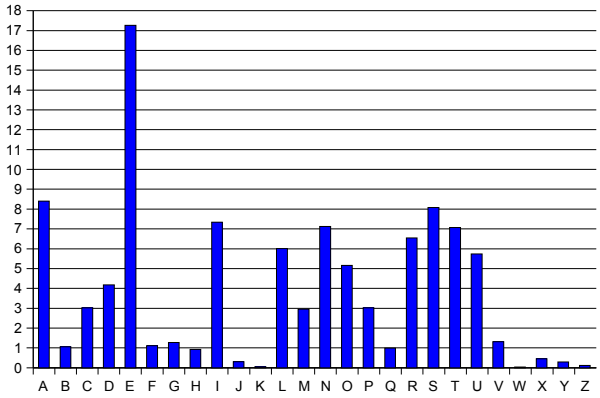
Après avoir constaté que la moyenne de classe était catastrophique, le professeur décide de monter tout le monde d'un demi-point. Laquelle de ces mesures statistiques ne changera pas : la moyenne, l'écart-type, le mode ou la médiane ?

Exercice 1.12

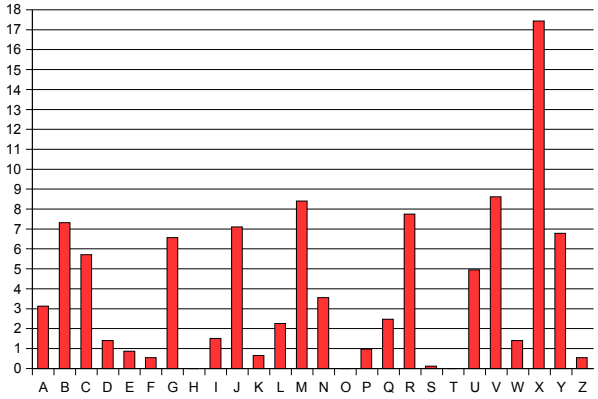
Déchiffrez le texte ci-dessous, sachant que chaque lettre du code remplace toujours la même lettre du texte original, écrit en français.

Un des moyens les plus simples de chiffrer un message est de remplacer chaque lettre par une autre. Ce chiffre a bien résisté aux cryptanalystes, jusqu'à ce que le savant arabe **Abu Yusuf Ya'qub ibn Is-haq ibn as-Sabbah Oòmran ibn Ismaïl al-Kindi** mette au point, au 9e siècle, une technique dite **analyse des fréquences** : comme chaque symbole correspond à une seule lettre, les fréquences d'apparition doivent être semblables. Ainsi, la lettre « e » est la plus utilisée en français, donc la lettre qui la remplace dans le message codé doit l'être aussi. Cependant, cette technique ne marche que si le message chiffré est assez long pour avoir des moyennes significatives.

XY AXJ BYRJMYJ, MQQMVUVXYJ GXR NCBWJR N'UYX LMBY N'PCLLX XJ BGR XAVBDBVXYJ, XY IMAX NU AMYNXGMFVX, RUV GX QGMJVX NU LUV NU QMGMBR VCEMG. GX VCB DBJ AXJJX QMVJBX NX LMBY KUB XAVBDMBJ. MGCVR GX VCB APMYWXM NX ACUGXUV, RXR QXYRXXR G'XIIVMEXVXYJ, GXR SCBYJUVXR NX RXR VXBYR RX NXGBXVXYJ XJ RXR WXYCUZ RX PXUVJXVXYJ G'UY G'MUJVX. GX VCB AVBM MDXA ICVAX QCUV IMBVX DXYBV GXR LMWBABXYR, GXR APMGNXXYR XJ GXR MRJVCGCWUXR. GX VCB QVBJ GM QMVCGX XJ NBJ MUZ RMWXR NX FMFEGCYX : JCUJ PCLLX KUB GBVM AXJJX XAVBJUVX XJ LX IXVM ACYYMBJVX RCY XZQGBAMJBCY VXDXJBVM GM QCUVQVX, LXJJVM GX ACGGBXV N'CV M RCY ACU XJ, ACLLX JVCBRBXLX NMYR GX VCEMULX, BG ACLLMYNXVM. MGCVR DBYVXYJ JCUR GXR RMWXR NU VCB, LMBR BGR YX QUVXYJ QMR GBVX G'XAVBJUVX XJ IMBVX ACYYMBJVX MU VCB G'XZQGBAMJBCY. GX VCB FMGJPMRMV IUJ NCYA JVXR XIIVMEX, GM ACUGXUV NX RCY DBRMWX APMYWXM XJ RXR WVMYNR IUVXYJ FCUGXDXVRXR. GM VXBYX, XY VMBRCY NXR QMVCGXR NU VCB XJ NX RXR WVMYNR, DBYJ NMYR GM RMGGX NU IXRJBY. GM VXBYX QVBJ GM QMVCGX XJ NBJ : KUX GX VCB DBDX XJXVYXGGXLXYJ ! KUX JXR QXYRXXR YX J'XIIVMEXYJ QMR XJ KUX JCY DBRMWX YX APMYWX QMR NX ACUGXUV. BG E M NMYR JCY VCEMULX UY PCLLX KUB QCRRXNX XY GUB G'XRQVBJ NXR NBXUZ RMBYJR.



Fréquences théoriques des lettres en français



Fréquences des lettres du cryptogramme

1.5. D'autres moyennes

A côté de la moyenne arithmétique que nous avons vue dans ce cours, il existe d'autres moyennes.

Moyenne géométrique

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Notation : $\prod_{i=1}^{N} x_1 + x_2 + \ldots + x_N$

On peut l'illustrer avec le cas suivant : si l'inflation d'un pays est de 5% la première année et de 15% la suivante, l'augmentation moyenne des prix se calcule grâce à la moyenne géométrique des coefficients multiplicateurs 1,05 et 1,15 soit une augmentation moyenne de 9,88%.

Exercice 1.13

On suppose qu'à l'issue d'une manifestation, la police annonce 10'000 manifestants, et les organisateurs 100'000. Quel est le nombre de manifestants ?
On se dit que les organisateurs et la police trichent de la même façon : si x est le nombre de manifestants réel, alors, si les organisateurs annoncent k fois plus de manifestants, la police en annonce k fois moins.

Moyenne harmonique

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

Si un train fait un trajet aller-retour entre deux villes à la vitesse moyenne v_1 pour l'aller et à la vitesse moyenne v_2 au retour, la vitesse moyenne du trajet complet n'est pas la moyenne arithmétique des deux vitesses, mais bien leur moyenne harmonique.

Exercice 1.14

Un avion a fait le trajet de A vers B, contre le vent, à la vitesse moyenne de 700 km/h et le trajet retour à 900 km/h. Quelle a été sa vitesse moyenne ?

Moyenne quadratique

$$\bar{x} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$$

Si un rectangle a pour côtés 3 et 7, le carré qui a même diagonale que le rectangle a pour côté la moyenne quadratique de 3 et 7, c'est-à-dire 5.38.

Moyenne pondérée

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

Le p_i sont les poids de chaque valeur.

Exemple : un prof qui donne différents poids à ses épreuves utilisera la moyenne pondérée.

1.6. Ce qu'il faut absolument savoir

Dessiner un histogramme	❑ ok
Dessiner un diagramme à secteurs	❑ ok
Dessiner un polygone des fréquences cumulées	❑ ok
Calculer une moyenne, un écart-type, une médiane, un intervalle semi-interquartile et un mode dans le cas discret	❑ ok
Calculer une moyenne, un écart-type, une médiane, un intervalle semi-interquartile et un mode dans le cas continu.	❑ ok
Connaître les différentes moyennes	❑ ok

Gotlib, La Rubrique-à-Brac, tome 1, page 85, Dargaud, 1970

En complément de ce chapitre, vous trouverez des exercices avec un tableur sur la page :

www.apprendre-en-ligne.net/madimu/tableur/

2. Ajustements

2.1. Un peu d'histoire

Adrien-Marie **Legendre**
(Paris, 18/9/1752 -
Paris, 10/1/1833)

Le problème de l'ajustement d'un ensemble de points représentés dans un système d'axes par une droite, ou plus généralement par une courbe, est essentiel dans le développement de la statistique.
Au 18ème siècle, Leonhard **Euler** et Tobias **Mayer** développent, indépendamment l'un de l'autre, la méthode des moyennes permettant d'ajuster des points par une droite.
Le premier texte paru faisant mention de la méthode des moindres carrés est dû à Adrien-Marie **Legendre** dans un article sur ses « nouvelles méthodes pour la détermination des orbites des comètes », publié en 1805. Un an plus tard, **Gauss** fait aussi allusion à cette méthode. C'est avec l'apparition de la loi normale que cette méthode va trouver sa justification et va devenir pour longtemps LA méthode d'ajustement.
La paternité de la corrélation a donné lieu à une littérature abondante. Signalons simplement que **Galton** exprime le désir de construire un coefficient de réversion qui se mutera en régression et qu'en 1888 il utilise les termes de « partial co-relation » annonçant déjà la corrélation multiple. En 1896, Karl **Pearson** reprend les concepts de **Galton** pour leur donner leur forme actuelle. Au 20ème siècle, d'autres mesures d'association allaient naître comme, en 1904, le coefficient de corrélation de rang avec **Spearman** et la même année la statistique « classique » du chi-deux par **Pearson**.

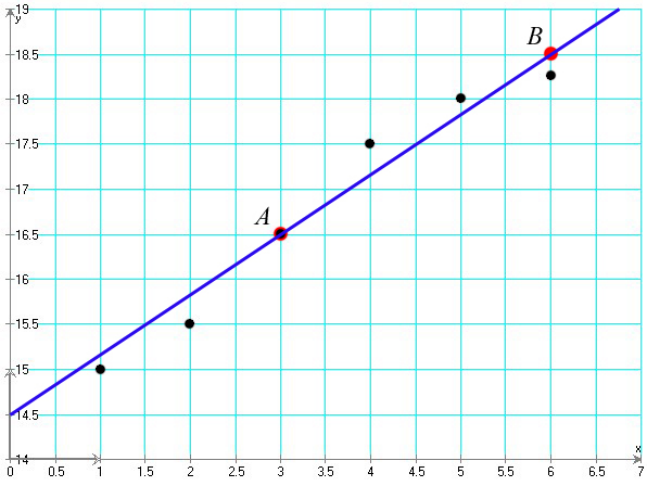
2.2. Ajustement affine graphique

Soient les n points du nuage représentant, dans un repère cartésien, la série des n valeurs (x_i, y_i) des variables x et y. Ajuster une droite d à ce nuage de points consiste à remplacer chaque point (x_i, y_i) par un point de même abscisse et d'ordonnée $\hat{y}_i$, les points $(x_i, \hat{y}_i)$ étant alignés sur la droite d.

Une fois l'équation de la droite d déterminée, on pourra l'utiliser pour faire des *interpolations* (calculs de valeurs intermédiaires) et des *extrapolations* (calculs de valeurs futures).

La méthode graphique consiste à tracer, **à l'œil**, à l'aide d'une règle **transparente**, une droite $y = m \cdot x + h$ s'ajustant le mieux possible sur le nuage de points.

Les points noirs représentent les données.
Les points rouges A et B sont les points choisis pour tracer la droite. Ils peuvent être choisis parmi les points noirs (A) ou pas (B).



Équation de la droite d'ajustement

Une fois la droite tracée, on choisit sur le dessin deux points A et B quelconques de la droite pour en déterminer l'équation. Ces points ne doivent pas obligatoirement faire partie du nuage de points.

Cette méthode est couramment employée, en raison de sa rapidité et de sa simplicité. Elle est empirique, mais donne de très bons résultats.

empirique : basé sur l'expérience

Rappel : L'équation de la droite passant par les points $A(x_A, y_A)$ et $B(x_B, y_B)$ est donnée par : $y - y_B = \frac{y_B - y_A}{x_B - x_A}(x - x_B)$

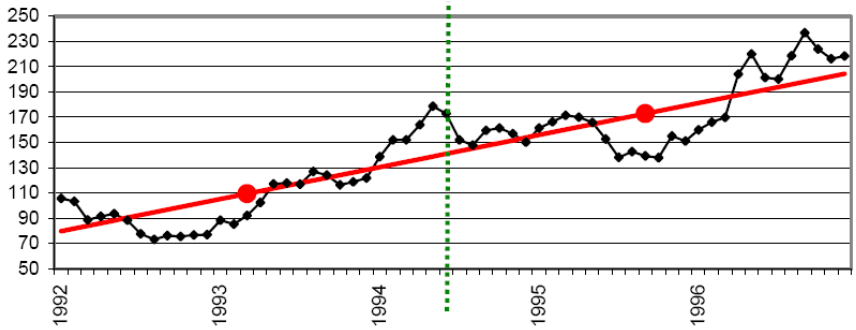
Les points A et B choisis dans notre exemple ont comme coordonnées (3, 16.5) et (6, 18.5). La droite passant par ces deux points est :

$$y - 18.5 = \frac{18.5 - 16.5}{6 - 3}(x - 6)$$

On obtient après simplification : $y = \frac{2}{3}x + 14.5$.

2.3. Droite de Mayer

Méthode On découpe le nuage de points en deux sous-ensembles de même effectif. Pour chacun des deux sous-ensembles, on calcule la moyenne des x_i et la moyenne des y_i. On obtient ainsi deux points $(\bar{x}_1, \bar{y}_1)$ et $(\bar{x}_2, \bar{y}_2)$, appelés **points moyens**. Il reste à tracer la droite passant par ces deux points.



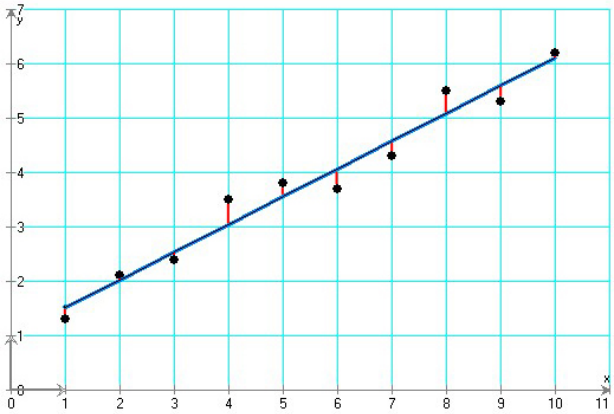
L'équation de cette droite s'obtient de la même façon que pour un ajustement affine graphique.

2.4. Ajustement analytique par la méthode des moindres carrés

$\hat{y}_i$ est la coordonnée verticale du point **de la droite** d'abscisse x_i. Donc $\hat{y}_i = a\,x_i + b$.

L'ajustement linéaire par la méthode des moindres carrés consiste à déterminer la droite (que l'on appelle aussi **droite de régression**) telle que la somme des carrés des n valeurs $y_i - \hat{y}_i$ soit minimale (ce qui explique le nom de la méthode).

Sur le dessin, chaque trait vertical rouge représente la valeur $y_i - \hat{y}_i$



On veut donc minimiser la quantité $q=\sum (y_i-(a\,x_i+b))^2$.

Karl Pearson
(Londres, 27/3/1857 -
Coldharbour, 27/4/1936)

Rappelons que la valeur minimale d'une fonction se calcule en posant sa dérivée égale à 0. Pour trouver *a* et *b*, calculons cette dérivée.

Calculons d'abord la dérivée de *q* par rapport à *a*.

$$\frac{dq}{da}=-2\sum ((y_i-a\,x_i-b)\,x_i)=0$$

$$\sum x_i\,y_i=a\sum x_i^2+b\sum x_i \qquad (1)$$

Calculons maintenant la dérivée de *q* par rapport à *b*.

$$\frac{dq}{db}=-2\sum (y_i-a\,x_i-b)=0$$

$$\sum y_i=\sum a\,x_i+\sum b$$

$$\sum y_i=\sum a\,x_i+nb \qquad \text{Divisons le tout par } n$$

$$\bar{y}=a\,\bar{x}+b$$
$$b=\bar{y}-a\,\bar{x} \qquad (2)$$

Ce résultat indique que la droite passe par le point moyen $(\bar{x};\bar{y})$.
Introduisons le résultat de (2) dans (1) pour trouver *a* :

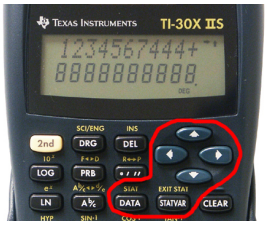
$$\sum x_i\,y_i=a\sum x_i^2+(\bar{y}-a\,\bar{x})\sum x_i$$

$$\sum x_i\,y_i=a\sum x_i^2+\bar{y}\sum x_i-a\,\bar{x}\sum x_i$$

$$a\sum x_i^2-a\,\bar{x}\sum x_i=\sum x_i\,y_i-\bar{y}\sum x_i$$

$$a=\frac{\sum x_i\,y_i-\bar{y}\sum x_i}{\sum x_i^2-\bar{x}\sum x_i}$$

$$a=\frac{\sum x_i\,y_i-n\,\bar{x}\,\bar{y}}{\sum x_i^2-n\,\bar{x}^2}=\frac{\frac{1}{n}\sum x_i\,y_i-\bar{x}\,\bar{y}}{\frac{1}{n}\sum x_i^2-\bar{x}^2}=\frac{\sigma_{xy}}{\sigma_x^{\,2}}$$



La droite des moindres carrés $y = ax + b$ a pour coefficients :

$$a=\frac{\frac{1}{n}\sum x_i\,y_i-\bar{x}\,\bar{y}}{\frac{1}{n}\sum x_i^2-\bar{x}^2} \qquad \text{et} \qquad b=\bar{y}-a\,\bar{x}$$

Remarque Certaines calculatrices ont des fonctions statistiques qui fournissent ces valeurs très rapidement. Consultez le mode d'emploi de votre machine !

Exercice 2.1

Lors d'une expérience, on a relevé les valeurs suivantes :

x	1	2	3	4	5	6	7	8	9	10
y	1.1	3.1	4.7	7.3	9.2	11.1	12.9	15.4	17	18.8

Donnez l'équation d'une droite ajustant ces valeurs

a. à l'œil ;
b. par la méthode de Mayer ;
c. par la méthode des moindres carrés.
d. Dessinez les droites obtenues en **b** et **c.**
e. Interpolez la valeur de $\hat{y}$ pour $x = 6.3$ grâce aux droites obtenues en **b** et **c.**

Exercice 2.2

Le tableau ci-dessous compare des voitures de même catégorie. Il met en rapport la cylindrée (en pouces^3) et le nombre de miles parcourus avec un gallon d'essence (3.78 litres aux USA).

Voiture	Cylindrée	Miles par gallon
VW Rabbit	97	24
Datsun 210	85	29
Chevette	98	26
Dodge Omni	105	24
Mazda 626	120	24
Oldsmobile Starfire	151	22
Mercury Capri	140	23
Toyota Celica	134	23
Datsun 810	146	21

Donnez l'équation d'une droite ajustant ces valeurs

a. à l'œil ;
b. par la méthode des moindres carrés.
c. Dessinez la droite obtenue en **b.**
d. Estimez le nombre de miles par gallon d'une voiture ayant une cylindrée de 125 grâce à la droite obtenue en **b.**

Exercice 2.3

Le tableau ci-dessous montre l'évolution des temps olympiques du 200 m plat, en secondes, pour les hommes et pour les femmes.

Vous remarquerez que les mesures au centième de seconde apparaissent en 1968 pour les hommes et en 1972 pour les femmes.

	200 m hommes	200 m femmes
Londres 1948	21.1	24.4
Helsinki 1952	20.7	23.7
Melbourne 1956	20.6	23.4
Rome 1960	20.5	24.0
Tokyo 1964	20.3	23.0
Mexico 1968	19.83	22.5
Munich 1972	20.00	22.40
Montréal 1976	20.23	22.37
Moscou 1980	20.19	22.03
Los Angeles 1984	19.80	21.81
Séoul 1988	19.75	21.34
Barcelone 1992	19.73	21.72
Atlanta 1996	19.32	22.12
Sydney 2000	20.09	21.84
Athènes 2004		

Donnez l'équation des droites (celle des performances des hommes et celle des performances des femmes) ajustant ces valeurs

a. à l'œil ;
b. par la méthode des moindres carrés.
c. Dessinez les droites obtenues en **b**.
d. Estimez les temps olympiques de 2004.
e. D'après les droites obtenues en **b**, en quelle année les femmes courront-elles le 200 m plat aussi vite que les hommes ?
f. Ces ajustements affines sont-ils adéquats ?

2.4. Coefficient de corrélation linéaire

Définition On nomme **coefficient de corrélation linéaire** des variables x et y, le nombre réel :

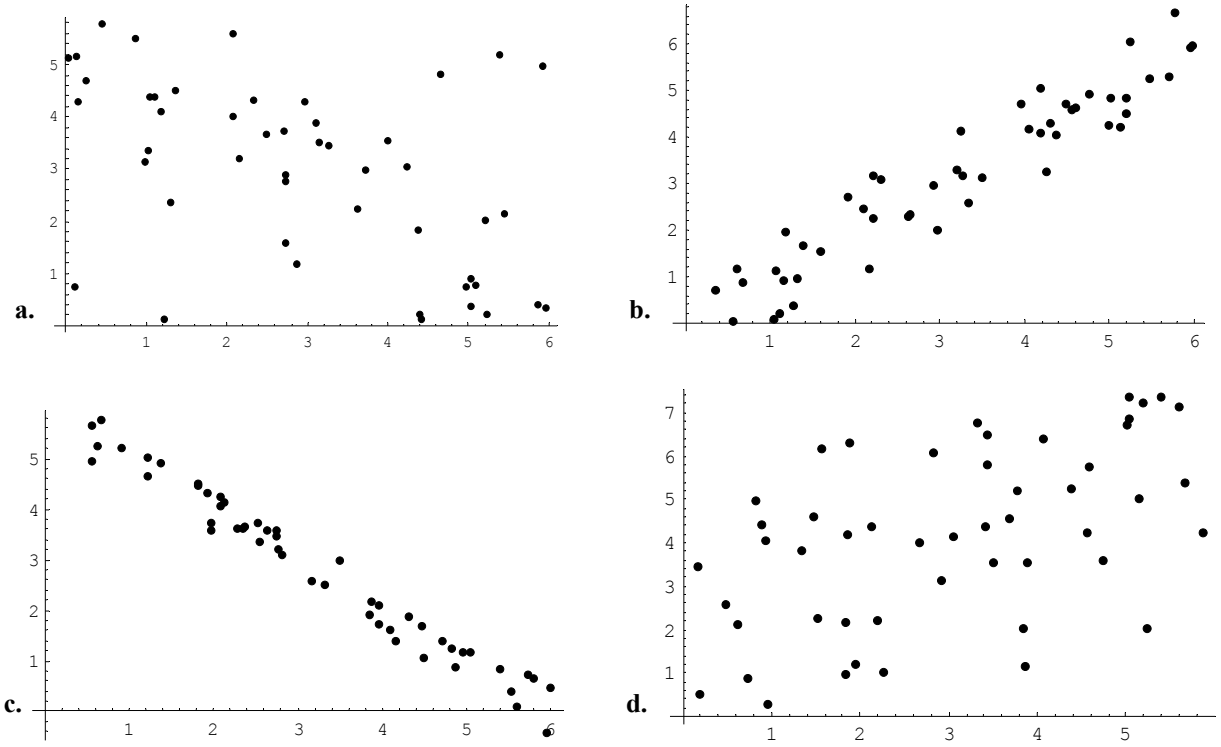
$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

avec $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$ $\quad \sigma_x = \sqrt{\frac{1}{n}\sum_{1}^{n} x_i^2 - \bar{x}^2}$ $\quad \sigma_y = \sqrt{\frac{1}{n}\sum_{1}^{n} y_i^2 - \bar{y}^2}$

Interprétation r est un nombre réel compris entre –1 et 1.
Quand | r | = 1, tous les points sont alignés.
Quand | r | est proche de 1, les variables x et y sont fortement corrélées.
Quand $r < 0$, la droite de régression a une pente négative.
Quand $r > 0$, la droite de régression a une pente positive.

Exercice 2.4

Rendez à chacun des nuages de points ci-dessous son coefficient de corrélation linéaire : –0.98, –0.50, 0.53, 0.94.



Exercice 2.5

Les criquets ont un organe spécial sur leurs ailes avant qui produit un son lorsqu'ils frottent leurs ailes les unes contre les autres. En règle générale, plus la température de l'air est élevée, plus ils frottent leurs ailes rapidement. La relation entre la température et le nombre de pulsations par seconde est bien approchée par une droite de régression (chaque espèce a sa droite propre). On a relevé les mesures suivantes :

Température (°C) [x]	15°	17°	20°	21°	23°	24°	27°	28°	30°	32°	34°
# de pulsations par sec. [y]	13.5	14.1	14.5	14.4	16.3	15.5	17.1	17.8	18.2	20.2	20.1

a. Donnez la droite des moindres carrés ajustant ce nuage de points.
b. Calculez le coefficient de corrélation linéaire.
c. Si la température augmente de 3°C, de combien augmentera le nombre de pulsations ?

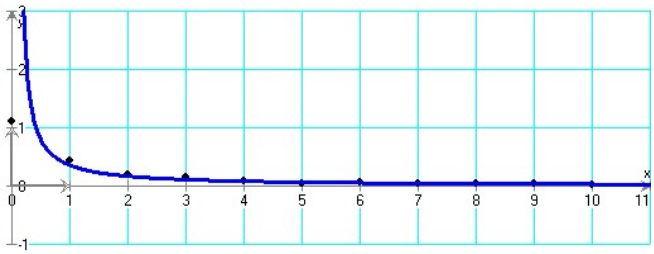
2.5. Autres ajustements

Ajustement par une hyperbole

Les points $(x_i\,;y_i)$ ne sont pas alignés, mais plutôt proches d'une certaine hyperbole de la forme $\hat{y}=\dfrac{1}{a\,x+b}$.

droite de régression de z en x : z est la valeur à estimer et joue le rôle de l'ordonnée, x joue le rôle de l'abscisse.

1. calculer $z_i=\dfrac{1}{y_i}$;
2. déterminer l'équation de la droite de régression de z en x avec la méthode des moindres carrés ;
3. de l'équation obtenue $z=a\,x+b$, on déduit immédiatement l'équation de l'hyperbole $\hat{y}=\dfrac{1}{a\,x+b}$.



Exercice 2.6

Ajustez ce nuage de points par une hyperbole $\hat{y}=\dfrac{1}{a\,x+b}$.

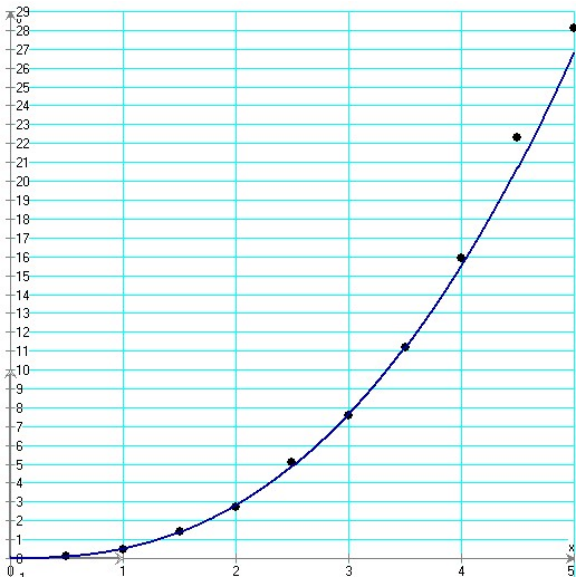
x	0	1	2	3	4	5	6	7	8	9	10
y	1.1	0.43	0.19	0.15	0.08	0.05	0.06	0.05	0.04	0.04	0.03

Ajustement par une fonction puissance

Les points $(x_i\,;y_i)$ sont proches d'une courbe de fonction puissance comme $\hat{y}=b\,x^a$. On remarque que $\ln(y) = a\ln(x) + \ln(b)$.

droite de régression de v en u : v est la valeur à estimer et joue le rôle de l'ordonnée, u joue le rôle de l'abscisse.

1. calculer $u_i=\ln(x_i)$ et $v_i=\ln(y_i)$;
2. déterminer l'équation de la droite de régression de v en u avec la méthode des moindres carrés ;
3. de l'équation obtenue $v = Au + B$, on déduit l'équation de la fonction puissance $\hat{y}=b\,x^a$, puisque $a = A$ et $b = e^B$.



Exercice 2.7

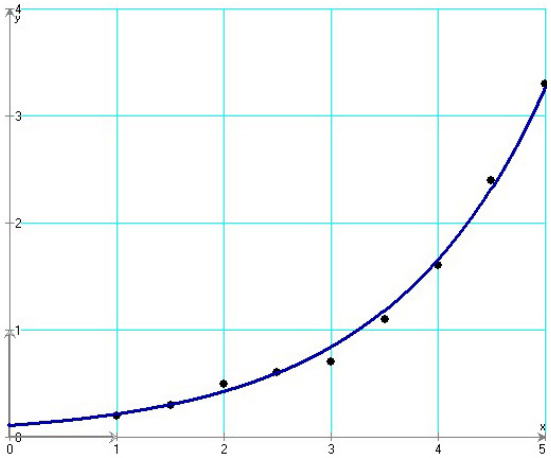
Ajustez ce nuage de points par une fonction puissance $\hat{y}=b\,x^{a}$.

x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	0.1	0.5	1.4	2.7	5.1	7.6	11.2	15.9	22.3	28.1

Ajustement par une exponentielle

Les points $(x_i\,;\,y_i)$ sont proches d'une courbe d'une exponentielle de la forme $\hat{y}=b\,a^{x}$.
On remarque que $\ln(y) = x\ln(a) + \ln(b)$.

1. calculer $z_i=\ln(y_i)$;
2. déterminer l'équation de la droite de régression de z en x avec la méthode des moindres carrés ;
3. de l'équation obtenue $z = Ax + B$, on déduit l'équation de l'exponentielle $\hat{y}=b\,a^{x}$, puisque $a = e^{A}$ et $b = e^{B}$.



Exercice 2.8

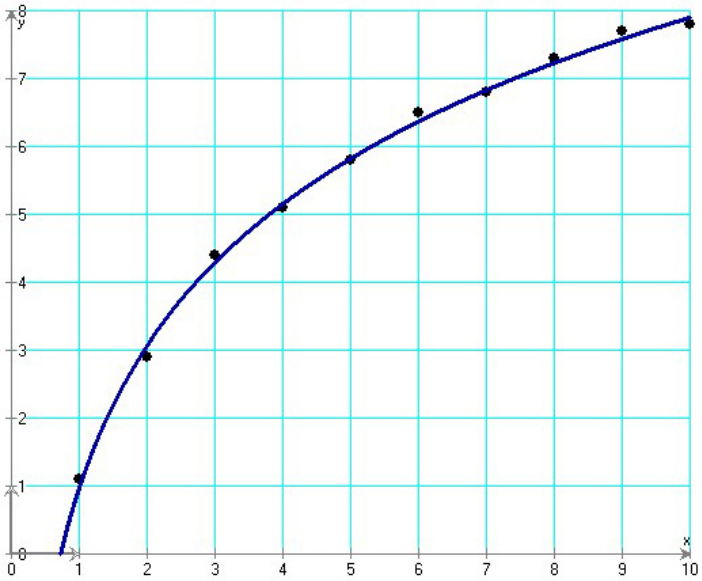
Ajustez ce nuage de points par une exponentielle de la forme $\hat{y}=b\,a^{x}$.

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	0.2	0.3	0.5	0.6	0.7	1.1	1.6	2.4	3.3

Ajustement par une fonction logarithmique

Les points $(x_i\,;y_i)$ sont proches d'une courbe logarithmique de la forme $\hat{y}=a\ln(x)+b$.

1. calculer $z_i=\ln(x_i)$;
2. déterminer l'équation de la droite de régression de y en z avec la méthode des moindres carrés ;
3. de l'équation obtenue $y = a{\cdot}z + b$, on déduit l'équation de la fonction logarithmique $\hat{y}=a\ln(x)+b$.



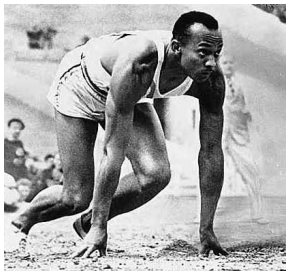
Exercice 2.9

Ajustez ce nuage de points par une fonction logarithmique $\hat{y}=a\ln(x)+b$.

x	1	2	3	4	5	6	7	8	9	10
y	1.1	2.9	4.4	5.1	5.8	6.5	6.8	7.3	7.7	7.8

Exercice 2.10

On veut étudier l'évolution des records de l'épreuve d'athlétisme du 100 mètres masculin.
Pour cela, on cherche un ajustement des records pour en prévoir l'évolution. On donne dans le tableau suivant certains records, établis depuis 1900.



Année	1900	1912	1921	1930	1964	1983	1991	1999
Rang (x_i)	0	12	21	30	64	83	91	99
Temps en sec. (y_i)	10.80	10.60	10.40	10.30	10.06	9.93	9.86	9.79

1) Étude d'un modèle affine

a. Construisez le nuage de points $M(x_i\,;y_i)$, avec i compris entre 1 et 8, associé à cette série statistique double. Vous prendrez comme unité graphique 1 cm pour dix ans en abscisse et 1 cm pour un dixième de seconde en ordonnées.
On commencera les graduations au point de coordonnées (0 ; 9).

b. Peut-on envisager un ajustement affine à court terme ? Cet ajustement permet-il des prévisions pertinentes à long terme sur les records futurs ?

2) Étude d'un modèle exponentiel

Après étude, on choisit de modéliser la situation par une autre courbe. On effectue les changements de variables suivants : $X = e^{-0.00924x}$ et $Y = \ln y$. On obtient le tableau :

$X = e^{-0.00924x}$	1.000	0.895	0.824	0.758	0.554	0.464	0.431	0.401
$Y = \ln y$	2.380	2.361	2.342	2.332	2.309	2.296	2.288	2.281

a. Donnez une équation de la droite de régression de Y en X obtenue par la méthode des moindres carrés.

b. En déduire que l'on peut modéliser une expression de y en fonction de x sous la forme suivante :

$y = \exp(a \cdot e^{-0.00924x} + b)$ où a et b sont deux réels à déterminer.

c. A l'aide de cet ajustement, quel record du 100 mètres peut-on prévoir en 2010 ?

$\exp(x) = e^x$

d. Calculez la limite en $+\infty$ de la fonction f définie sur $\mathbb{R}$ par l'expression suivante :

$f(x) = \exp(0.154\ e^{-0.00924x} + 2.221)$

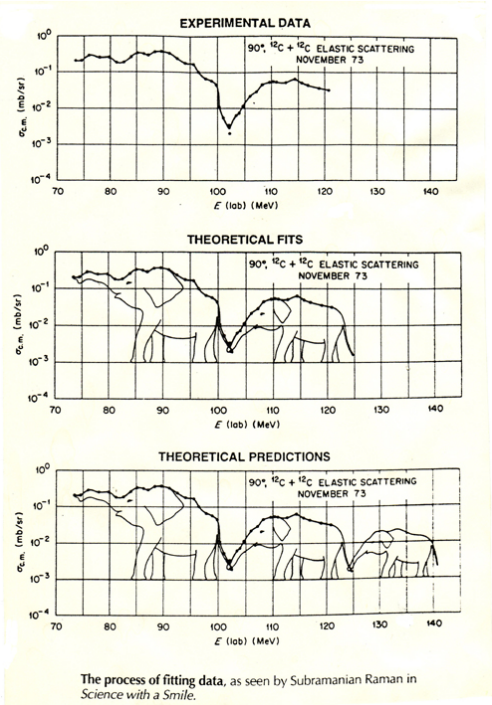
e. Que peut-on en conclure, en utilisant ce modèle, quant aux records du cent mètres masculin, à très long terme ?

2.6. Ce qu'il faut absolument savoir

Faire un ajustement affine graphique	❑ ok
Faire un ajustement affine par la méthode de Mayer	❑ ok
Faire un ajustement affine par la méthode des moindres carrés	❑ ok
Estimer et interpréter un coefficient de corrélation linéaire	❑ ok
Faire un ajustement par une hyperbole	❑ ok
Faire un ajustement par une fonction puissance	❑ ok
Faire un ajustement par une exponentielle	❑ ok
Faire un ajustement par une fonction logarithmique	❑ ok

En complément de ce chapitre, vous trouverez des exercices avec un tableur sur la page :

www.apprendre-en-ligne.net/madimu/tableur/



The process of fitting data, as seen by Subramanian Raman in *Science with a Smile.*

Solutions des exercices

Chapitre 1

1.1. **a.** 10 **b.** 24 **c.** 20 **d.** 54

1.2. **a.** 0 **b.** 2 **c.** 1.581

1.3. $\bar{x}=23.55$, $\tilde{x}=20.5$, mode = 19,
$v = 79.74$, $\sigma = 8.93$, isi = 2

1.4. $\bar{x} = 9.822$, $\sigma = 0.222$

1.5. 5.6

1.6. environ 3'173'000

1.8. **c.** mode = 123.64, $\tilde{x}=122.22$,
$isi=\frac{130.91-112.14}{2}=9.385$

1.9. **c.** mode $\cong$ 3475 francs, $\tilde{x}=3525$,
$isi=\frac{Q_3-Q_1}{2}\approx\frac{3810-3290}{2}=260$
d. $\bar{x} \cong 3559$, $\sigma = 393.6$

1.10. **a.** $\bar{x} = 55.370$ **b.** 68.52 % des écoles

1.11. **c1.** $\bar{x} = 10.938$, $\sigma = 0.530$
c2. $\bar{x} = 10.944$, $\sigma = 0.545$
d. non

1.13. 31600 personnes

1.14. 787.5 km/h

Chapitre 2

2.1. **b.** $\hat{y} = 1.99x - 0.9$
c. $\hat{y} = 1.992x - 0.896$
e. 11.64 et 11.65

2.2. **b.** $\hat{y} = -0.0837x + 34.0092$
d. 23.5467

2.3. **b.** hommes : $\hat{y} = 66.336 - 0.0234x$
femmes : $\hat{y} = 122.173 - 0.0504x$
d. hommes : 19.44 (en réalité : 19.79)
femmes : 21.17 (en réalité : 22.05)
e. en 2068

2.4. **a.** –0.98 **b.** –0.50
c. 0.53 **d.** 0.94

2.5. **a.** $\hat{y} = 7.3 + 0.37x$
b. 0.97
c. 1.11

2.6. $\hat{y}=\frac{1}{3.1x-0.33}$

2.7. $\hat{y}=0.52x^{2.45}$

2.8. $\hat{y}=0.11\cdot 1.97^{x}$

2.9. $\hat{y} = 3\ln(x) + 0.99$

2.10. **1) b.** oui pour le court terme (coefficient de corrélation linéaire proche de 1) ; non pour le long terme (au bout d'un certain temps, le temps serait nul).
2)
a. $Y = 0.154X + 2.221$
b. $\ln(y)=0.154e^{-0.00924x}+2.221$
c. 9.74
d. 9.21
e. selon ce modèle, l'homme ne pourra pas courir le 100 m en moins de 9.21 secondes.