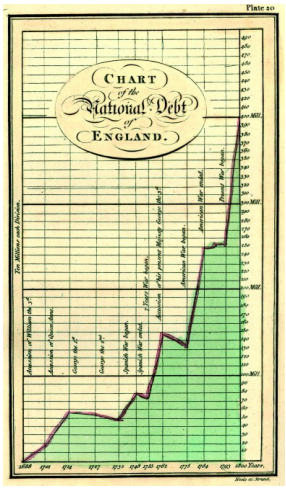
1. Statistique descriptive

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 19$^{\text{ème}}$ 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, un 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 : $n=\sum_{i=1}^{7} n_i=26$

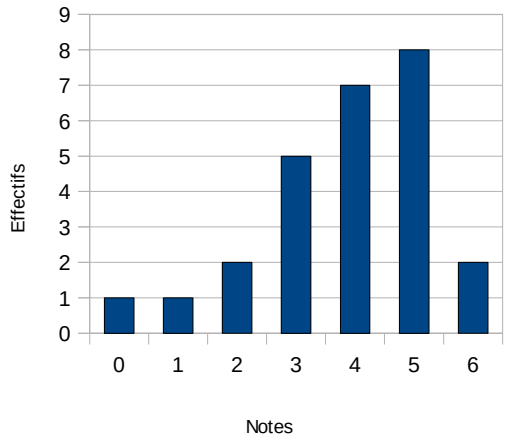
Notation : $\sum_{i=1}^{N} x_i = x_1 + x_2 + \ldots + x_N$

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.** $\sum_{i=1}^{4} n_i \cdot \sum_{j=1}^{4} x_j$

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 de gauche ci-dessous sont dessinés d'après les données présentées dans le tableau 1.



Histogramme

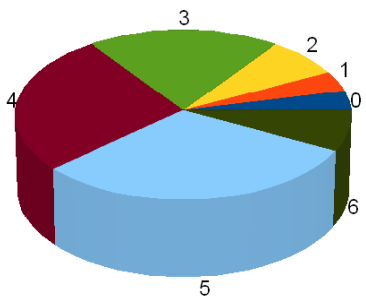
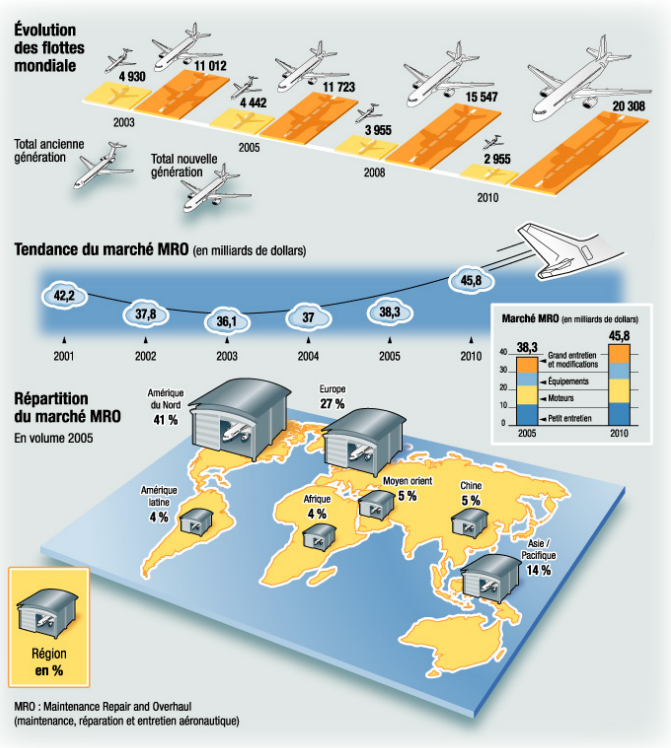


Diagramme à secteurs



Infographie de presse réalisée pour le magazine interne d'Air France (source : www.seguier.fr)

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.

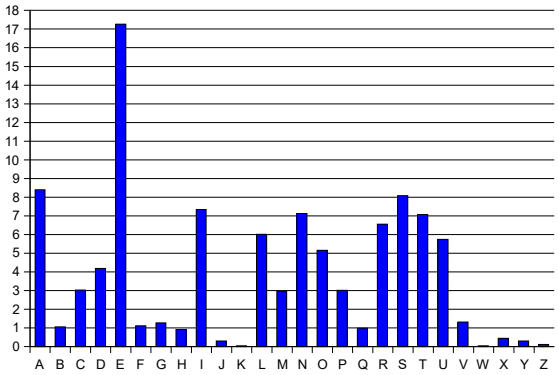
Jouez le jeu ! Les résultats seront rassemblés et analysés en classe.

Exercice 1.8

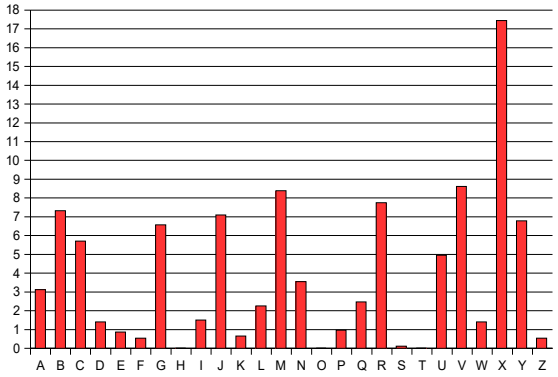
Déchiffrez le texte ci-dessous, sachant que chaque lettre du code remplace toujours la même lettre du texte original, écrit en français.

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

1.4. 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
		$n = 40$

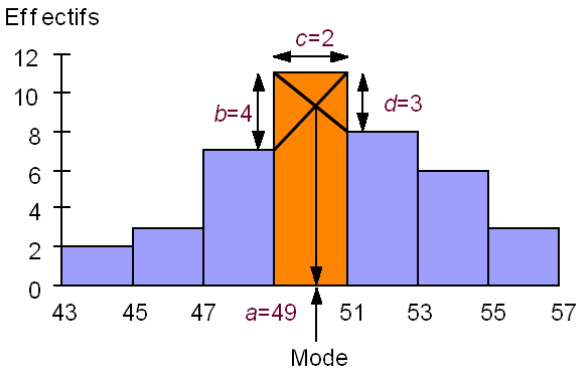
Mode

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 $x_i \leq x$.

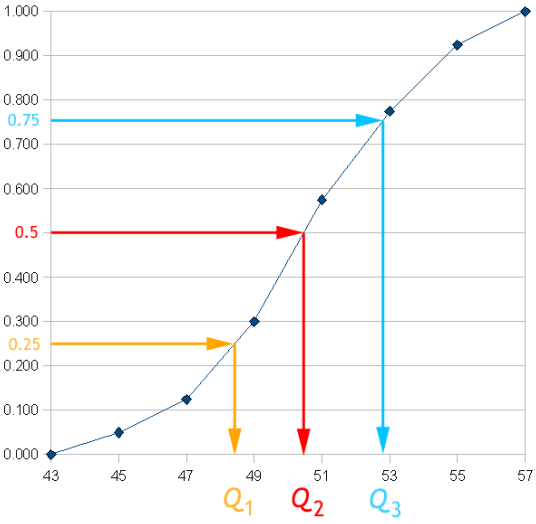
Ce tableau représente les vitesses de 40 voitures mesurées dans un village.

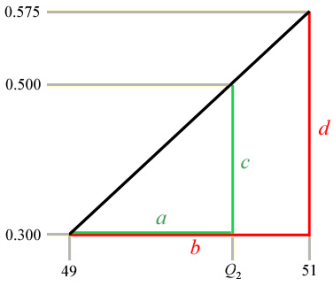
Tableau 3

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
		$\sum = 40$	$\sum = 1$	

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*).

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

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 1.9

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 1.10

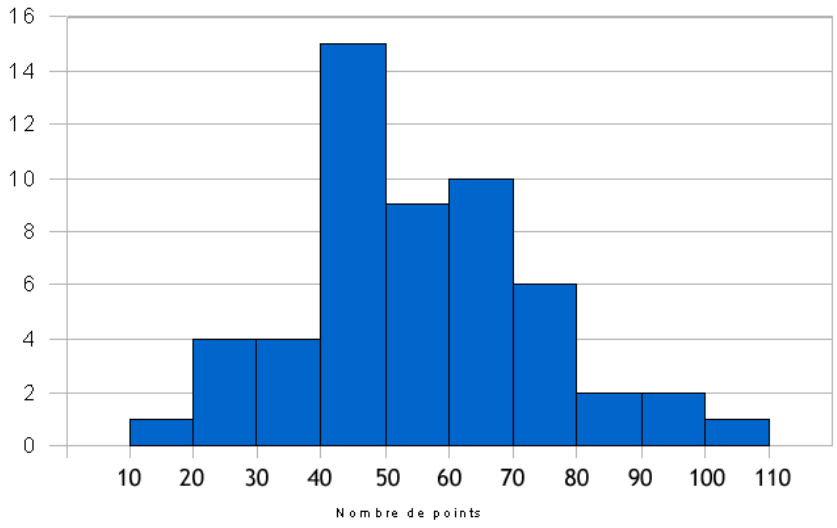
Les salaires mensuels payés aux ouvriers d'une entreprise se répartissent comme suit :

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 1.11

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



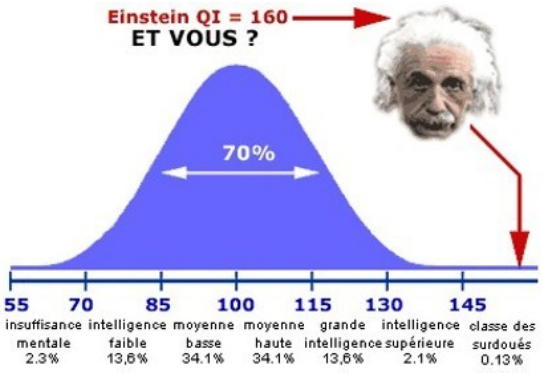
a. Calculez la moyenne de cette série.
b. En utilisant l'histogramme, trouvez le pourcentage des écoles qui ont moins de 64 points.

Exercice 1.12

Après avoir constaté que la moyenne de classe était (une fois de plus) catastrophique, le prof de maths 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 1.13

Créé au début du 20e siècle pour dépister les élèves en difficulté et leur faire bénéficier d'un soutien, le test de QI est très vite détourné à des fins eugénistes pour isoler et formater certains enfants supposés avoir le meilleur potentiel. Si le test était très algébrique il permettrait des scores extrêmes : score très élevé pour un enfant « normalement intelligent » mais souffrant d'autisme ou score faible pour un enfant « normalement intelligent » mais souffrant de dyslexie. Dès lors, les tests comprennent des questions « culturelles » ne mesurant plus la capacité de calcul, mais l'érudition et l'apport des parents.



Un groupe de psychologues a élaboré un test qui attribue à chaque personne un nombre Q qui mesure ses capacités intellectuelles (plus Q est grand, plus les capacités sont élevées).
Supposons que chacun des habitants de deux pays A et B ait obtenu son nombre Q. On prend alors pour le niveau intellectuel de chaque pays la moyenne arithmétique des nombres Q de ses habitants.

Un groupe d'habitants du pays A a émigré dans le pays B.
a. Est-il possible que le niveau intellectuel des deux pays ait augmenté ?

Après cela, un groupe d'habitants du pays B (parmi lesquels il peut y avoir des anciens émigrés de A), émigre dans le pays A.
b. Est-il possible que le niveau intellectuel des deux pays augmente de nouveau ?

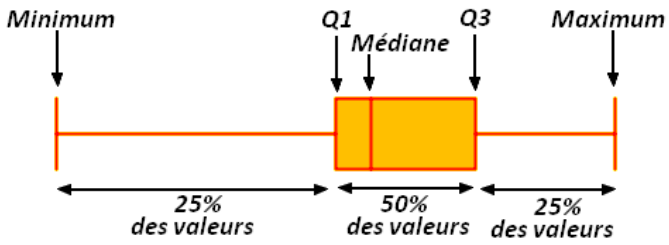
Un groupe d'habitants a émigré du pays A dans le pays B et un autre groupe du pays B dans le pays C. Cela a fait augmenter le niveau intellectuel de chacun des trois pays. Ensuite les flots migratoires ont changé de direction : un groupe d'habitants a émigré de C dans B et un autre de B dans A. Les agences d'information des trois pays affirment que le niveau intellectuel de chaque pays a augmenté encore plus après cette deuxième migration.
c. Est-ce que c'est possible ?

On suppose qu'entre les migrations le quotient intellectuel Q de chaque personne ne change pas, qu'il n'y a eu aucune naissance et aucun décès.

1.5. Boîte à moustaches


John Wilder **Tukey**
(1915-2000)

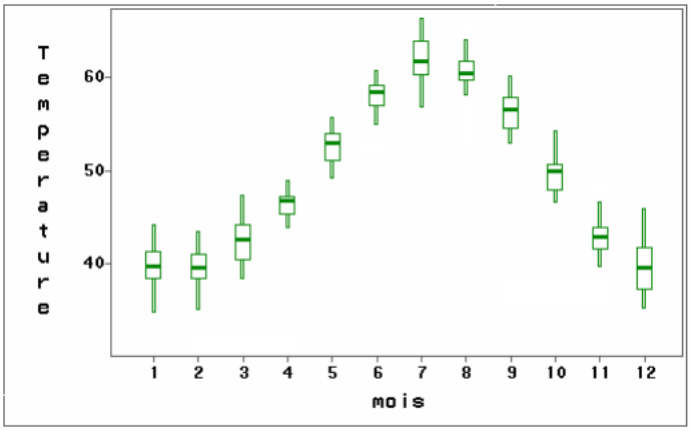
Dans les représentations graphiques de données statistiques, la boîte à moustaches ou diagramme en boîte ou encore diagramme à pattes est un moyen rapide de figurer le profil essentiel d'une série statistique. Elle a été inventée en 1977 par John **Tukey**, mais peut faire l'objet de certains aménagements. Son nom est la traduction de *Box and Whiskers Plot*.
La boîte à moustaches résume certaines caractéristiques de position du caractère étudié (médiane, quartiles, minimum, maximum ou déciles). Ce diagramme est utilisé par exemple pour comparer un même caractère dans deux populations de tailles différentes.



Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. Ce rectangle suffit pour le diagramme en boîte. On ajoute alors des segments aux extrémités menant jusqu'aux valeurs extrêmes.
Ces boîtes à moustaches peuvent aussi être dessinées verticalement.

Exemple

Soit la série des températures mensuelles moyennes à Nottingham de 1920 à 1939. Les 240 données proviennent de : http://robjhyndman.com/tsdldata/data/anderson15.dat.
Ces données ont été regroupées par mois et représentées sous forme de boîtes à moustaches :



1.6. D'autres moyennes

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

Moyenne géométrique

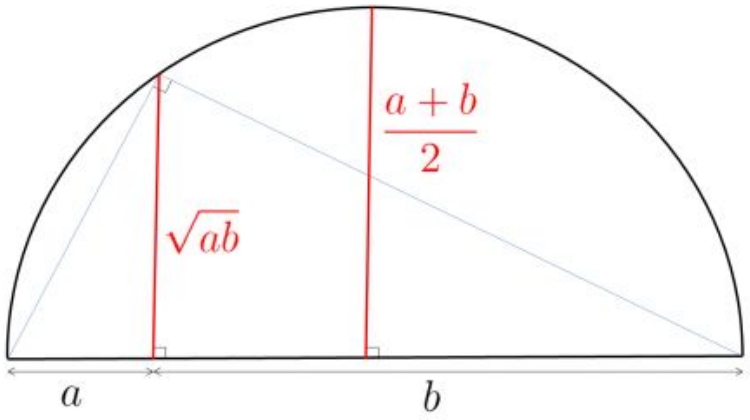
$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Notation :

$$\prod_{i=1}^{N} x_1 \cdot x_2 \cdot \ldots \cdot x_N$$

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%.

La moyenne géométrique est toujours inférieure ou égale à la moyenne arithmétique, comme le montre la preuve sans mot ci-dessous :



La moyenne géométrique en politique

Le 30 mars 2014, Philippe Perrenoud a été réélu grâce à ce particularisme bernois qui prévoit que la minorité du Jura bernois a droit à un siège au Conseil d'État, et qu'il est occupé par le candidat qui obtient la meilleure moyenne géométrique.
Cette moyenne s'obtient en prenant la racine carrée de la multiplication des scores que les candidats francophones obtiennent dans l'ensemble du canton et dans le Jura bernois.

Cette spécificité du système électoral a été imaginée suite à l'éviction en 1986 de Geneviève Aubry. Porté par les électeurs alémaniques, un inconnu avait taillé une vexante veste à la radicale qui avait pourtant réalisé le double de voix dans le Jura bernois. Elle n'avait jamais posé question jusqu'ici.

Candidats :	Votes obtenus dans tout le canton	Votes obtenus dans le Jura bernois	Résultat de la moyenne géométrique
Philippe Perrenoud	86'469	5'889	22'566
Manfred Bühler	94'957	4'919	21'612

Grâce à cette formule, Philippe Perrenoud est élu au Conseil-Exécutif malgré un score obtenu dans l'ensemble du canton de Berne qui est inférieur à celui obtenu par Manfred Bühler.
On voit que cette moyenne donne un poids élevé aux votes du Jura bernois. Avec la moyenne arithmétique, c'est Manfred Bühler qui aurait été élu...

Exercice 1.14

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

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

La moyenne harmonique intervient notamment dans le calcul de la longueur des cordes des instruments de musique (d'où son nom), car la fréquence produite est inversement proportionnelle à la longueur de la corde.
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 (voir exercice 1.15).

Exercice 1.15

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 ?

Exercice 1.16

On change 100 euros en dollars au taux de 0.70 euro pour un dollar et 100 euros au taux de 0.80 euro pour un dollar. Quel est le taux de change moyen ?

Moyenne quadratique

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

Elle est utilisée pour calculer l'écart-type (voir page 3).
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.

Les moyennes vues ci-dessus peuvent être ordonnées ainsi :

Harmonique < Géométrique < Arithmétique < Quadratique

Moyenne pondérée

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

Le p_i sont les poids de chaque valeur.
Un prof qui donne différents poids à ses épreuves utilise la moyenne pondérée.

1.7. Paradoxe de Simpson

Ce paradoxe porte le nom d'Edward Simpson qui l'a étudié pour la première fois.

En 1973, l'université américaine de Berkeley (Californie), fut poursuivie pour discrimination envers les filles. L'affaire semblait claire : parmi les candidates, seules 30 % étaient retenues, alors que 46 % des candidatures masculines l'étaient.
L'étude a été précisée sur les six départements les plus importants, notés ici de A à F :

Département	Garçons	Admis	Filles	Admises
A	825	62 %	108	82 %
B	560	63 %	25	68 %
C	325	37 %	593	34 %
D	417	33 %	375	35 %
E	191	28 %	393	24 %
F	272	6 %	341	7 %
Total	2590	46 %	1835	30 %


Edward **Simpson** (né en 1922)

Ce tableau, si l'on excepte la dernière ligne, ne montre aucune discrimination envers les femmes. Au contraire, le taux d'admission des filles dans le principal département (A) est nettement supérieur à celui des garçons. L'explication de ce paradoxe apparent vient quand on regarde le nombre de candidatures dans ces départements. Les femmes semblent avoir tendance à postuler en masse à des départements très sélectifs. Dans ceux-ci, leur taux d'admission est à peine plus faible que celui des hommes. Dans les autres, elles sont plus largement sélectionnées que les hommes. Quand on fait la moyenne globale, ce sont les départements sélectifs qui ont plus de poids, puisqu'elles y postulent en masse.

1.8. 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

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

www.nymphomath.ch/madimu/tableur/