Chapitre 4 : Statistiques et Échantillonnage

CONTENUS	CAPACITÉS ATTENDUES	COMMENTAIRES
Statistique descriptive, analyse de données Caractéristiques de position et de dispersion • médiane, quartiles ; • moyenne.	• Utiliser un logiciel (par exemple, un tableur) ou une calculatrice pour étudier une série statistique. • Passer des effectifs aux fréquences, calculer les caractéristiques d'une série définie par effectifs ou fréquences. • Calculer des effectifs cumulés, des fréquences cumulées. • Représenter une série statistique graphiquement (nuage de points, histogramme, courbe des fréquences cumulées).	L'objectif est de faire réfléchir les élèves sur des données réelles, riches et variées (issues, par exemple, d'un fichier mis à disposition par l'INSEE), synthétiser l'information et proposer des représentations pertinentes.
Échantillonnage Notion d'échantillon. Intervalle de fluctuation d'une fréquence au seuil de 95%*. Réalisation d'une simulation.	• Concevoir, mettre en œuvre et exploiter des simulations de situations concrètes à l'aide du tableur ou d'une calculatrice. • Exploiter et faire une analyse critique d'un résultat d'échantillonnage.	Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience. À l'occasion de la mise en place d'une simulation, on peut : • utiliser les fonctions logiques d'un tableur ou d'une calculatrice, ◇ mettre en place des instructions conditionnelles dans un algorithme. L'objectif est d'amener les élèves à un questionnement lors des activités suivantes : • l'estimation d'une proportion inconnue à partir d'un échantillon ; • la prise de décision à partir d'un échantillon.

I] Statistiques

a) Rappels sur le calcul de la moyenne :

Définition : la moyenne d'une série statistique est égale à la somme des valeurs individuelles divisée par le nombre total de valeurs.

Exemples : Si vous avez eu 10, 12, 15 et 8 en maths au 1er trimestre, votre moyenne est égale à (10+12+15+8)÷4 soit 45÷4 = 11,25. Parfois, comme il y a plusieurs valeurs identiques, on comptabilise les valeurs en utilisant des *effectifs* (nombre de valeurs identiques) comme dans l'exemple suivant donnant les résultats des élèves d'une classe d'un contrôle :

Note obtenue	9	10	11	12	13	14	15	16	17	18	19	20	Total
Effectifs	2	4	4	6	7	4	3	5	2	0	2	1	40

Il y a eu 40 notes et le total des notes obtenues est 538 car :
9×2+10×4+11×4+12×6+13×7+14×4+15×3+16×5+17×2+19×2+20×1 = 538.
Donc la moyenne de la classe est 538÷40 = 13,45.
La moyenne sans tenir compte des effectifs (9+10+11+12+13+14+15+16+17+19+20)÷11 ≈ 14,18 n'a aucune signification ici (sauf d'être la moyenne des différentes notes obtenues par les élèves), en tout cas ce n'est pas la moyenne de la classe (des 40 notes).

La formule que l'on retient pour calculer $\overline{x}$, la moyenne de n valeurs x_i pondérées par les effectifs n_i (i est ici une variable muette qui prend toutes les valeurs de 1 à n) est la suivante : $\overline{x}=\frac{n_1 x_1+n_2 x_2+n_3 x_3+\ldots n_k x_k}{n_1+n_2+n_3+\ldots n_k}$.

Regroupement des individus en classes :
Jusqu'à présent, les grandeurs statistiques étudiées étaient *discrètes* (elles ne pouvaient prendre qu'un nombre fini de valeurs). Lorsque les grandeurs sont *continues* (une infinité de valeurs sont possibles), il n'est plus possible de compter l'effectif de chaque valeur. On est alors amené à regrouper les valeurs à l'intérieur de *classes* (groupe de valeurs proches). Dans certains cas, la variable est discrète mais il y a trop de valeurs pour les séparer toutes, on a alors aussi recours au regroupement par classe.

Exemple : On donne les effectifs et salaires des employés d'une Petite et Moyenne Entreprises (PME) :

Catégorie	Ouvrier simple	Ouvrier qualifié	Cadre moyen	Cadre supérieur	Dirigeant
Effectif	50	25	15	10	2
Salaire (en €)	de 800 à 1100	de 1100 à 1500	de 1500 à 2500	de 2500 à 5000	de 5000 à 11000
S_i : Salaire moyen (en €)	950	1300	2000	3750	8000

Ici, on a estimé le salaire moyen des employés d'une catégorie (classe socio-économique) en prenant les « milieux » des classes (on parle alors plutôt de *centres de classe*). Le salaire moyen dans une classe est une valeur arbitraire, que peut-être aucun employé ne va toucher. La moyenne des salaires pour l'entreprise se calculera alors en effectuant le quotient du total des salaires (estimé en additionnant les produits des effectifs par les valeurs moyennes S_i d'une classe) par l'effectif total :
Salaire moyen dans la PME $\overline{s}=\frac{50\times950+25\times1300+15\times2000+10\times3750+2\times8000}{50+25+15+10+2}=\frac{163500}{102}\approx1602{,}94$ €.

La formule que l'on retient ici est celle qui permet de calculer les centres $\overline{x_i}$ d'une classe $[a_i\,;\,b_i[$, c'est-à-dire la moyenne entre les bornes de l'intervalle : $x_i=\frac{a_i+b_i}{2}$. On est en droit de s'interroger sur la validité de cette formule pour certains cas, où la variable est discrète, comme dans notre exemple. L'intervalle « de 800 à 1100 € » s'écrit-il [800 ; 1100[? En principe oui, et l'on comptera dedans les salaires allant jusqu'à 1099,99 € ce qui fait qu'on peut bien compter 1100 comme borne supérieure de l'intervalle dans le calcul du centre de classe : $x_1=\frac{800+1100}{2}=\frac{1900}{2}=950$ €.

Acquisition des données : Les données sont parfois fournies par un graphique comme un *histogramme* ou un *diagramme circulaire* qui correspondent finalement à des tableaux d'effectifs ou de fréquences déguisés par la représentation en rectangles (histogrammes) ou en secteurs angulaires. Elles sont parfois fournies de façon brute, par la liste de toutes les valeurs obtenues.
Exemples : On donne une série de 40 mesures du diamètre de tubes PVC fabriqués par une usine :

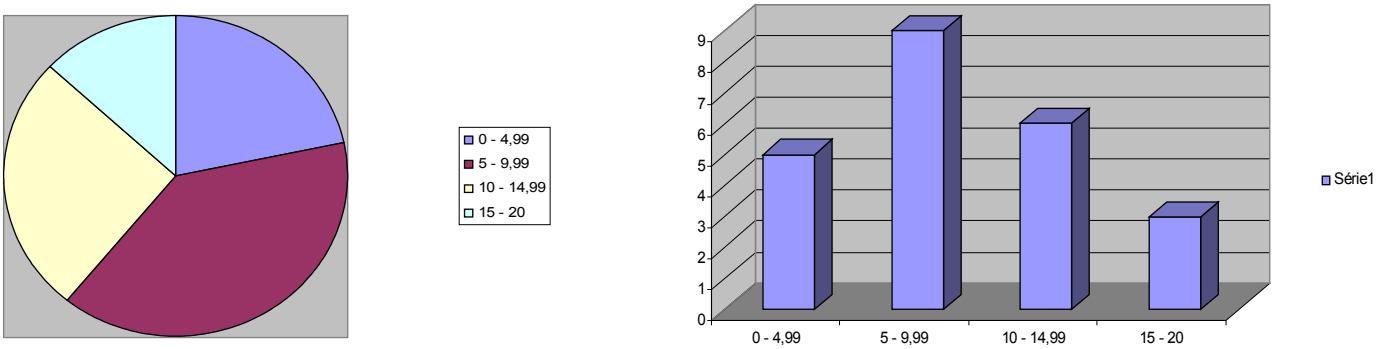
12,5 – 12,4 – 12,5 – 12,4 – 12,5 – 12,5 – 12,6 – 12,4 – 12,5 – 12,5
12,5 – 12,6 – 12,7 – 12,5 – 12,5 – 12,6 – 12,6 – 12,5 – 12,4 – 12,4
12,5 – 12,5 – 12,5 – 12,6 – 12,5 – 12,4 – 12,6 – 12,4 – 12,5 – 12,4
12,5 – 12,4 – 12,5 – 12,5 – 12,5 – 12,5 – 12,4 – 12,5 – 12,4 – 12,5.

Le premier travail est de construire le tableau des effectifs, en comptant méthodiquement les valeurs identiques (la méthodologie est importante ici pour des séries comportant beaucoup de valeurs).

Valeurs (D_i) en mm	12,4	12,5	12,6	12,7	total
Effectifs (n_i)	11	22	6	1	40

La moyenne de cette série est simple à calculer, maintenant qu'on a compté les effectifs :
$\overline{D}=\frac{12{,}4\times11+12{,}5\times22+12{,}6\times6+12{,}7\times1}{11+22+6+1}=\frac{499{,}7}{40}\approx12{,}4925$ mm.

Voici un autre exemple où le tableau des effectifs se déduit de la lecture d'un graphique : on donne la répartition du nombre d'élèves exclus quotidiennement des cours d'un collège sur une période de 23 jours.

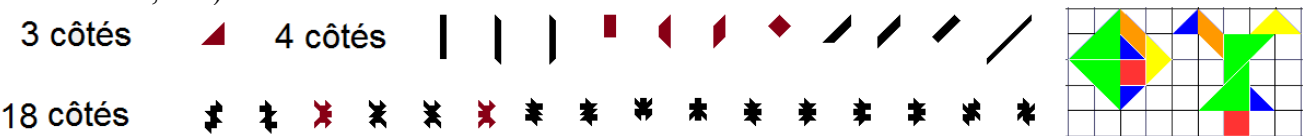


De l'un ou l'autre de ces graphiques, on doit être en mesure de tirer l'information qui se résume dans le tableau des effectifs suivant, duquel on déduit la formule de calcul de la moyenne :

Exclus de cours (par jour)	de 0 à 5	de 5 à 10	de 10 à 15	de 15 à 20
Effectifs	5	9	6	3
Valeur centrale	2,5	7,5	12,5	17,5

Nombre moyen d'élèves exclus = (2,50×5+7,5×9+12,5×6+17,5×3) ÷ (5+9+6+3) = 207,5 ÷ 23 ≈ 9,02.
NB : Les centres de classes sont calculés comme précédemment indiqué, mais on peut remarquer que cette façon de faire revient à calculer $x_1=\frac{0+5}{2}=\frac{5}{2}=2,5$ alors qu'en réalité cette classe contient les nombres de l'ensemble $\{0;1;2;3;4\}$ et si l'on en fait la moyenne, cela donne $x_1=\frac{0+1+2+3+4}{5}=\frac{10}{5}=2$. Donc il convient mieux, dans le cas de variables discrètes, d'effectuer les calculs des centres de classe sur les extrémités réelles de l'intervalle et non sur des valeurs irréalistes.

Dans la plupart des cas réels, l'acquisition des données est réalisée par un procédé automatique plutôt que manuel. Les nombres sont écrits dans des fichiers sous une forme qui permet leur stockage, leur copie et leur utilisation ultérieure dans les traitements statistiques. L'utilisation des tableurs est, à ce titre très répandue, mais le tableur n'est pas le moyen le plus pratique pour effectuer des traitements sur de grosses quantités de données ou lorsque les données arrivent de façon continue (capteur de température, cours d'un titre en Bourse, etc.).



Exemple de calcul d'une moyenne avec un tableur : L'image ci-dessus montre quelques *polyabolos* (assemblages de triangles isocèles rectangles égaux obtenus en juxtaposant des côtés semblables) constitués de 16 triangles. Le tableau ci-dessous en donne la répartition selon le nombre de leurs côtés (ligne 2) qui va de 3 à 18. Sur l'image, nous avons colorié en rouge foncé les formes qui sont reconstituables avec les pièces du *Tangram* (comme on le voit à droite où le carré et une forme à 18 côtés sont reconstitués par les 7 pièces du célèbre puzzle). La ligne 19 du tableau, en dessous des effectifs (ligne 18), contient les produits n_ix_i nécessaires au calcul de la moyenne. Celle-ci est obtenue (case R19) en sommant ces nombres et en divisant le résultat obtenu par le total des effectifs (case R18).

R19 : =SOMME(B19:Q19)/R18

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		Côtés																
2	Triangles	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
18	16	1	13	31	325	1 982	10 937	47 554	175 058	528 695	1 326 781	2 726 852	4 485 875	5 689 037	5 276 394	3 210 730	982 983	24 463 248
19		3	52	155	1950	13874	87496	427986	1750580	5815645	15921372	35449076	62802250	85335555	84422304	54582410	17693694	14,89

Les moyens informatiques tels que le tableur évitent d'avoir à effectuer les calculs complexes et répétitifs avec tous les risques d'erreurs qui sont liés aux défauts de l'attention d'un opérateur humain. Cependant, ils requièrent, au moment de la conception des feuilles de calculs, toute l'attention du concepteur car le tableur n'exécute que ce qu'on lui demande (comme la calculatrice)...

Calcul algorithmique : À titre d'exemple on peut écrire l'algorithme qui permet de calculer la moyenne d'une série de données fournie par un fichier dont les enregistrements seraient écrits de la façon suivante :

- Chaque ligne contient deux données : la valeur (x_i) de la grandeur mesurée et l'effectif (n_i) de cette grandeur. Entre les deux nombres, il y a un séparateur, par exemple un espace (voir l'illustration ci-contre).
- Le fichier contient un certain nombre de lignes, non connu à l'avance, mais la fin du fichier est matérialisée par un indicateur, une ligne vierge par exemple.

```
3 1
4 13
5 31
6 325
7 1982
8 10937
9 47554
10 175058
11 528695
12 1326781
13 2726852
14 4485875
15 5689037
16 5276394
17 3210730
18 982983
```

L'algorithme doit lire les lignes, tester si la ligne n'est pas vierge et dans ce cas, lire les 2 nombres. Il doit effectuer les calculs nécessaires à la détermination de la moyenne et afficher celle-ci à la fin du traitement.

1. Déclaration des variables utilisées : XI, NI, T_X (total valeurs), T_N (total effectifs), I (indicateur lignes)
2. Initialisation des variables : XI=0, NI=0, T_X=0, T_N=0, I=0
3. Ouverture du fichier

4. Tant que la ligne n'est pas vierge :
   Lecture du 1er nombre→XI
   Lecture du 2ème nombre→NI
   Réaffectations : T_X= T_X+XI×NI ; T_N= T_N+NI ; I=I+1
5. Fermeture du fichier et affichage de la moyenne : (T_X÷T_N) et du nombre d'enregistrements : I.

b) Utilisation des fréquences

Les fréquences sont, en fait, des fractions calculées à partir des effectifs, par une de ces formules :

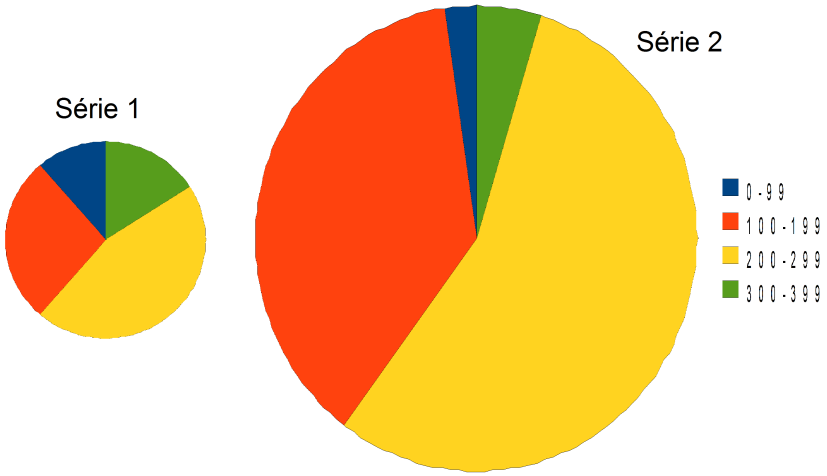
$$\text{Fréquence d'une classe (nombre décimal compris entre 0 et 1)} = \frac{\text{Effectif de la classe}}{\text{Effectif total}}$$

$$\text{Fréquence d'une classe (pourcentage compris entre 0 et 100 \%)} = \frac{\text{Effectif de la classe}}{\text{Effectif total}} \times 100\,\%$$

Les fréquences sont donc des nombres variant entre 0 et 1. Lorsqu'elles sont exprimées en pourcentage, on donne le numérateur d'une fraction dont le dénominateur est 100. Ces numérateurs varient donc entre 0 et 100, on les appelle les *taux* du pourcentage. Le premier intérêt des fréquences est de pouvoir comparer des séries qui ont des effectifs totaux différents. Ce qu'on compare alors, c'est la répartition des valeurs dans la série. Jugez plutôt sur cet exemple où l'on dispose de 2 séries de mesures faites par des observateurs différents :

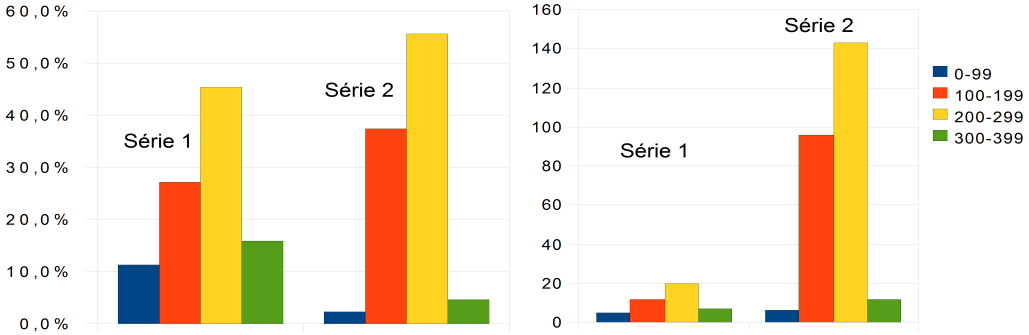
Valeurs	de 0 à 99	de 100 à 199	de 200 à 299	de 300 à 399	de 0 à 99	de 100 à 199	de 200 à 299	de 300 à 399
Effectifs	5	12	20	7	6	96	143	12
Fréquences	11,4%	27,3%	45,4%	15,9%	2,3%	37,4%	55,6%	4,7%

Afin d'effectuer les comparaisons plus facilement, sans avoir à lire les nombres, on a l'habitude de représenter la répartition des fréquences à l'aide d'un graphique (diagramme circulaire, rectangles, etc.). On peut choisir de représenter à la fois les différences de répartition et les différences d'effectifs par des diagrammes circulaires dont les aires sont proportionnelles aux effectifs totaux et les angles des secteurs proportionnels aux fréquences. Comme la série 2 a un effectif total presque 6 fois plus grand (257÷44 ≈ 5,84) que la série 1, il faut multiplier les 2 dimensions du diagramme (longueur et largeur) par $\sqrt{6} \approx 2{,}5$.



Lorsqu'il s'agit de grandeurs quantitatives, l'ordre des valeurs a du sens (on parle de grandeurs *ordinales*, qu'on peut ordonner, par opposition avec les grandeurs *catégoriques* ou *nominales*, qu'on ne peut pas ordonner). La représentation sous la forme d'un diagramme circulaire ne restitue pas l'ordre des valeurs. Cette représentation est mieux adaptée aux variables non-quantitatives (comme par exemple des catégories sociaux-professionnelles, des couleurs, etc.). Dans le cas des grandeurs ordinales, on pourra choisir des *histogrammes*.

Cela n'empêche pas de faire varier les dimensions des histogrammes si l'on veut rendre les différences d'effectifs totaux. Sinon on peut choisir de conserver les effectifs comme dans la représentation de droite ci-



dessus : la répartition entre les classe est restituée par les hauteurs de rectangles et les différences d'effectifs totaux apparaissent aussi naturellement. Dans la représentation de gauche, nous avons utilisé les fréquences. Il faudrait alors agrandir l'histogramme de la série 2 qui a un effectif presque 6 fois plus grand. Si on choisi d'agrandir juste les hauteurs (il suffit de les multiplier par 6), on se retrouve avec le graphique de droite (à une échelle près)!
Vous constatez à l'observation des histogrammes, qu'ici, les deux séries ont des profils assez différents : l'allure de la répartition des valeurs est assez « espacée » pour la première série (on dit "dispersée") dans le sens qu'il y relativement beaucoup de valeurs petites et grandes et peu de valeurs moyennes, alors qu'elle est assez « concentrée » pour la deuxième dans le sens qu'il y relativement peu de valeurs extrêmes (petites ou grandes) et beaucoup de valeurs moyennes. Cette caractéristique de plus ou moins grande concentration autour des valeurs centrales est qualifiée de *dispersion* de la série. Nous reviendrons plus loin sur les différentes méthodes d'estimation de la dispersion d'une série.

c) Calculs de la médiane et des quartiles

Pour étudier plus en détail une série statistique, on peut essayer de chercher la valeur qui partage la série en deux groupes de même effectif. C'est ce que l'on a appelé en 3ème la **médiane** de la série statistique (une valeur centrale de la série, généralement voisine de la moyenne mais rarement égale à celle-ci, car obtenue par un procédé différent). Le moyen le plus simple pour déterminer la médiane est de ranger toutes les valeurs dans l'ordre croissant et de déterminer celle qui est au centre.
Exemple : Les 40 mesures du diamètre de tubes PVC fabriqués par une usine données plus haut peuvent être ordonnées de façon croissante et comptabilisées au moyen des effectifs et effectifs cumulés. On peut dire qu'il y a 33 valeurs égales à 12,4 ou 12,5 ou inférieures à 12,6. La 20ème valeur (40÷2=20) est-elle au centre de la série ? Techniquement pas tout-à-fait, car il y en a 19 en dessous et 20 au dessus. Il faudrait déterminer une valeur entre la 20ème et la 21ème. Ici la 19ème et la 21ème sont égales à la 20ème, donc cela n'a pas vraiment d'importance. La médiane sera ici égale à 12,5. La moyenne, par comparaison, est légèrement différente puisqu'elle vaut 12,4925 (il n'y a une grande différence certes!).

Valeurs	12,4	12,5	12,6	12,7
Effectifs	11	22	6	1
Effectifs cumulés	11	33	39	40

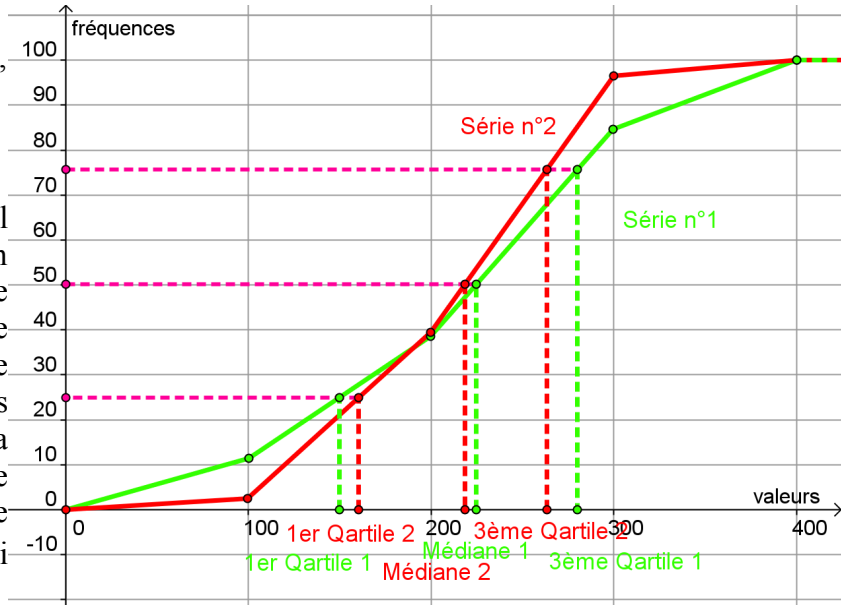
Malheureusement, ce n'est pas toujours possible de trouver une valeur exacte au centre de la série. On a souvent perdu les informations individuelles en regroupant les valeurs proches dans des classes (c'est toujours le cas lorsque la variable étudiée est continue). Dans ce cas, on va tout de même ordonner les classes et cumuler leurs effectifs (ou leurs fréquences). On peut ainsi déterminer la classe dans laquelle se situe la moitié de l'effectif (*classe médiane*), et au sein de cette classe, en supposant la répartition uniforme, faire ce qu'on appelle une interpolation linéaire pour déterminer la médiane. La médiane apparaît toujours, ainsi qu'elle est définie, comme la valeur correspondant à la moitié de l'effectif total (ou à la fréquence cumulée de 50%).

valeurs	de 0 à 100	de 100 à 200	de 200 à 300	de 300 à 400	de 0 à 100	de 100 à 200	de 200 à 300	de 300 à 400
effectifs	5	12	20	7	6	96	143	12
Effectifs cumulés	5	17	37	44	6	102	245	257
fréquences	11,4%	27,3%	45,4%	15,9%	2,3%	37,4%	55,6%	4,7%
Fréquences cumulées	11,4%	38,7%	84,1%	100%	2,3%	39,7%	95,3%	100%

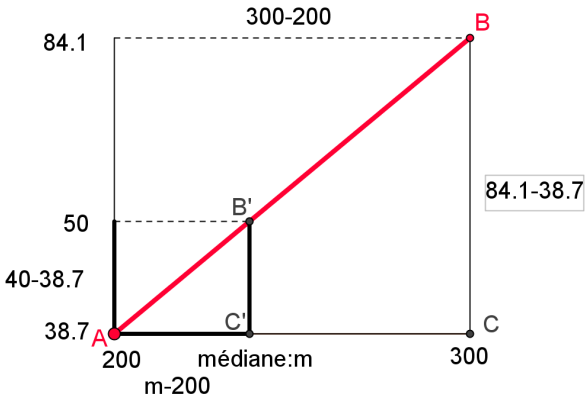
On peut lire les valeurs cumulées obtenues. Par exemple : 84,1% des valeurs de la 1ère série sont inférieures à 300. On utilise les fréquences car c'est plus parlant de faire ce type de lecture en fréquences plutôt qu'en effectifs (37 valeurs sur 44 sont inférieures à 300). Passons à l'estimation de la valeur médiane de la série.
Pour cela construisons la **courbe polygonale des fréquences cumulées** : Au point d'abscisse 100 on attribut l'ordonnée 11,4 (afin de pouvoir dire que 11,4% des valeurs sont inférieures à 100), etc. Lorsqu'on a placé ainsi 5 points (ne pas oublier celui d'abscisse 0 et d'ordonnée 0), on trace les 4 segments qui constituent cette ligne brisée appelée *polygone des fréquences cumulées en croissant* (on pourrait cumuler

en décroissant, c'est-à-dire en partant des valeurs supérieures). On peut maintenant *estimer* la médiane : en supposant que la répartition des valeurs dans chaque classe est homogène (on dit *uniforme*), c'est-à-dire qu'on peut les répartir selon les segments qui ont été tracés, il suffit de lire sur le graphique, en abscisse du point d'ordonnée 50%, la valeur de la médiane. Ainsi pour notre 1ère série, la médiane se situe environ à 220, tandis que pour la 2ème série la médiane est un peu inférieure. La moyenne, quant-à elle, se calcule toujours de la même façon (par la formule de la moyenne pondérée) :

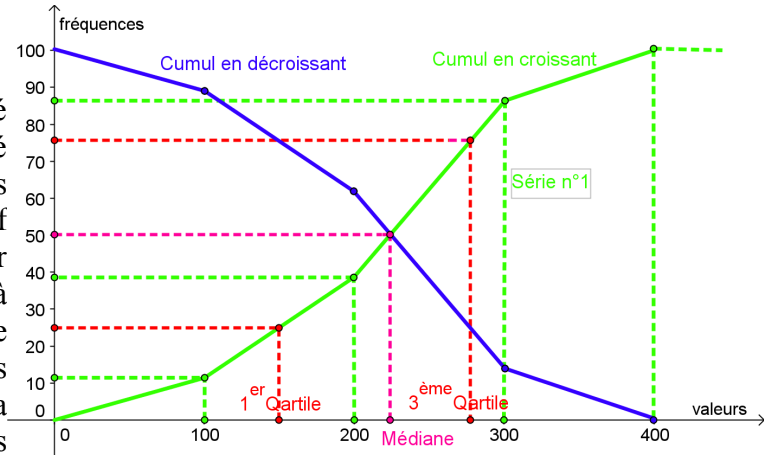
Avec les effectifs : moyenne 1 = $\frac{50\times5+150\times9+250\times20+350\times7}{44}\approx205$ ou bien,
Avec les fréquences : moyenne 1 = $\frac{50\times11{,}4+150\times27{,}3+250\times45{,}4+350\times15{,}9}{100}\approx205$.



Si l'on veut *calculer* la médiane, il faut traduire le fait que la répartition des valeurs dans la classe médiane suive le segment tracé. On se trouve dans une configuration du théorème de Thalès, celle des triangles emboîtés ABC et AB'C' (voir schéma ci-dessous), et il suffit alors d'écrire l'égalité des rapports qu'apporte ce théorème : $\frac{AC'}{AC}=\frac{B'C'}{BC}$ et donc ici $\frac{m-200}{300-200}=\frac{50-38{,}7}{84{,}1-38{,}7}$, ce qui conduit à : $m=200+(\frac{50-38{,}7}{84{,}1-38{,}7})\times300-200=200+(\frac{11{,}3}{45{,}4})\times100=200+(\frac{11{,}3}{45{,}4})\times100\approx224{,}8899$, environ 225 donc.



Notez qu'on peut, à l'aide d'une construction géométrique simple, déterminer cette valeur : il suffit de construire le *polygone des fréquences cumulées en décroissant*. Ce 2ème polygone coupe en effet le 1er en un point d'abscisse égale à la médiane! Cela n'a qu'une valeur anecdotique mais traçons ce 2ème polygone pour vérifier cette propriété.
En réalité, les deux polygones sont symétriques par rapport à l'axe horizontal qui passe par le point d'ordonnée 50%.



Deux autres paramètres statistiques ont été représentés sur le graphique sans avoir été définis : le 1er quartile et le 3ème quartile. Ces paramètres sont définis comme la médiane, sauf que le 1er quartile, noté parfois Q_1, est la valeur correspondant à un quart de l'effectif total (ou à 25% de la fréquence cumulée) et le 3ème quartile noté Q_3, est la valeur correspondant à trois quarts de l'effectif total (ou à 75% de la fréquence cumulée). La médiane est aussi, dans ce système de partage par quart, le 2ème quartile. On pourrait de la même manière définir des *déciles* (partage par dixièmes) par rapport auxquels la médiane serait le 5ème décile, ou des *centiles* (cela se pratique en statistiques descriptives).

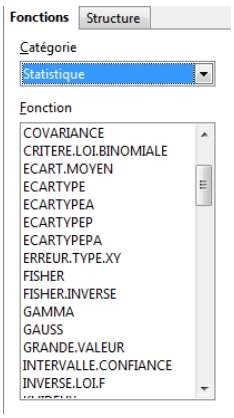
Remarque : On voit sur le graphique que le 1er et le 3ème quartiles sont plus proches dans le cas de la série n°2 que dans celui de la série n°1, ce qui traduit le fait que cette série est plus "concentrée", moins

"dispersée" que la 1ère série : les valeurs se regroupent davantage autour des valeurs centrales (médiane ou moyenne). Nous retrouvons ici la manifestation d'une notion déjà mentionnée plus haut (dans la partie *b,* sur les fréquences), celle de dispersion d'une série. Les paramètres Q_1 et Q_3 apparaissent ainsi comme une estimation de la dispersion de la série. Disons qu'en calculant leur différence, on tient un estimateur de cette dispersion. La *différence inter-quartile* est Q_3-Q_1, un *indice* qui mesure *la dispersion absolue*. Lorsqu'on rapporte cette différence à la médiane on obtient un *indice de dispersion relatif* qui peut être comparé à celui d'autres séries (ayant de médianes différentes).
Pour la 1ère série, on a $\frac{Q_3-Q_1}{Q_2}=\frac{280-150}{225}=\frac{130}{225}\approx 0{,}58$ alors que pour la 2ème on a $\frac{Q_3-Q_1}{Q_2}=\frac{265-160}{220}=\frac{105}{220}\approx 0{,}48$. Ces indices montrent bien, en le mesurant, que la 1ère série est moins concentrée que la 2ème. Plus l'écart est petit entre Q_1 et Q_3, et plus la dispersion est faible (les valeurs sont plus concentrées). Cet indicateur est proportionnel à la dispersion des valeurs, et inversement proportionnel à la concentration des valeurs.

On peut, en principe, estimer la dispersion d'une série relativement à une autre, à l'aide d'un autre paramètre, très simple à calculer : l'**étendue** de la série. L'étendue est égale à la différence entre la valeur maximum et la valeur minimum. Dans notre exemple, on n'obtient pas une grande information avec l'étendue car les valeurs ayant été regroupées en classe, on a perdu l'information utile, à savoir les valeurs extrêmes. On peut toujours dire qu'elles sont 0 et 400 pour les deux séries, mais c'est sans doute faux et cela n'apporte rien sur la connaissance de la dispersion. Le calcul de l'étendue pour les 40 mesures des diamètres de tubes PVC fabriqués par une usine (exemple de la partie *a*) donne 12,7─12,4 soit une étendue de 0,3. L'étendue est une mesure de la dispersion absolue. Rapportée à la moyenne, cela donne une estimation approximative de la dispersion relative, égale à 0,024 pour cet exemple. Une étendue relative très petite montre une dispersion très petite, et donc une concentration très élevée autour de la moyenne. L'étendue est sensible aux valeurs extrêmes (qui peuvent être aberrantes) alors que l'écart inter-quartile défini plus haut ne l'est pas. Cet écart mesure finalement l'étendue de la série après élimination de 25% des valeurs les plus faibles et de 25% des valeurs les plus élevées. On dit pour cela qu'il s'agit d'un indice plus robuste que l'étendue, la robustesse étant le contraire de la sensibilité, une insensibilité en quelques sortes, aux valeurs extrêmes.

d) Calculs de la moyenne des écarts et de l'écart-type

Un autre indicateur de la dispersion d'une série (ne figurant pas au programme) est la *moyenne des écarts* à la moyenne : on calcule tout d'abord la moyenne, puis on calcule les *écarts* (différence positive entre la plus grande et la plus petite valeur) entre les différentes valeurs et cette moyenne. Le calcul de la moyenne de ces écarts est donc une façon de mesurer l'éloignement moyen des valeurs par rapport à la moyenne, donc la dispersion : plus l'éloignement est grand, plus la dispersion est grande et inversement. Si l'on n'avait pas pris des écarts positifs (valeurs absolues des écarts) mais des écarts algébriques (positifs ou négatifs), nous aurions obtenu une moyenne nulle! Ceci vient de la définition même de la moyenne.
Encore un autre indicateur de la dispersion (ne figurant pas non plus au programme), est fourni par les calculatrices et très souvent employé dans tous les domaines recourant aux statistiques. Il s'agit de faire la moyenne des écarts aux carrés et d'en prendre la racine carrée. Cet indicateur s'appelle *écart-type*, et on le note avec la lettre grecque *sigma* : σ. L'intérêt de choisir les carrés pour rendre positifs les écarts algébriques ne nous apparaitra complètement que plus tard (la valeur absolue n'est pas dérivable alors que le carré l'est) mais soulignons que cela permet d'obtenir une formule qui simplifie le calcul de l'indicateur : alors que pour le calcul de la moyenne des écarts, il fallait effectuer tout d'abord le calcul de la moyenne et puis ensuite, calculer les écarts, ici on peut tout faire en un seul passage! On a en effet la relation suivante : σ^2=moyenne des carrés−carré de la moyenne. Pour calculer la moyenne on a dit qu'il fallait calculer les produits $n_i\ x_i$ et pour calculer la moyenne des carrés il faut calculer les produits $n_i\ x_i^2$ ce qui permet de tout faire en même temps. Cette remarque explique l'usage généralisé de σ, car on gagnait du temps à procéder ainsi avant l'arrivée des ordinateurs. Aujourd'hui, on peut facilement choisir un indicateur ou un autre, car les logiciels calculent tout cela sans aucune difficulté. Regardez dans votre tableur préféré, au menu des fonctions statistiques, il propose généralement tous ces indicateurs.

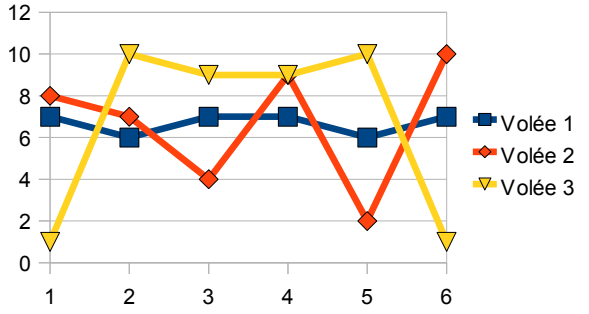


La moyenne des écarts, notée m_e, et l'écart-type σ sont des indicateurs de dispersion absolue. Si l'on veut mesure la dispersion relative (pour les comparer à ceux d'autres séries n'ayant pas les mêmes moyennes), il faut les diviser par la moyenne $\bar{x}$. On obtient ainsi $\frac{m_e}{\bar{x}}$, la *moyenne relative des écarts* et $\frac{\sigma}{\bar{x}}$ le *coefficient de variation*.

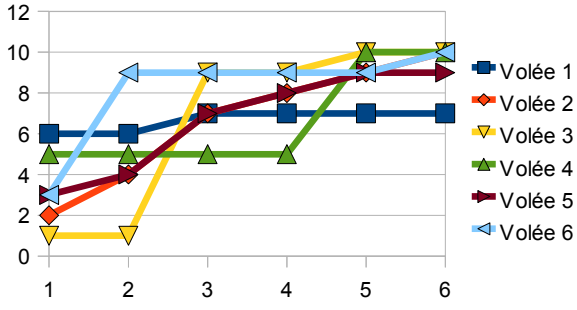
Exemples : Calculons les différents indicateurs absolus et relatifs de la dispersion que nous venons de définir pour deux séries différentes. Nous noterons q l'écart inter-quartile relatif égal à $\frac{Q_3-Q_1}{Q_2}$, $\frac{e}{\bar{x}}$ l'étendue relative, $\frac{m_e}{\bar{x}}$ la moyenne relative des écarts et $\frac{\sigma}{\bar{x}}$ le coefficient de variation. Ces deux séries sont choisies pour illustrer nos propos, elles ne sont donc pas tellement réalistes. Imaginons donc que des joueurs s'affrontent au tir à l'arc. Ils se placent à 70 m de la cible (122 cm de diamètre selon les prescriptions de la FFTA) dont les 10 anneaux concentriques ont les valeurs de 1 à 10 (une zone spéciale dans le centre sert à départager les meilleurs tireurs). Ils tirent chacun 6 volées de 6 flèches. Voici les résultats d'un tireur :

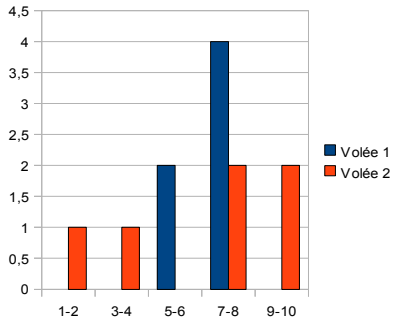
Flèche	1	2	3	4	5	6	Moyenne	Médiane	Écart-type	Écart-moyen	Distance interquartile	Étendue	Coefficient de variation	Distance interquartile relative
Volée 1	7	6	7	7	6	7	6,67	7	0,52	0,44	0,75	1	0,08	0,11
Volée 2	8	7	4	9	2	10	6,67	7,5	3,08	2,44	4	8	0,46	0,53
Volée 3	1	10	9	9	10	1	6,67	9	4,41	3,78	6,75	9	0,66	0,75
Volée 4	5	5	10	5	10	5	6,67	5	2,58	2,22	3,75	5	0,39	0,75
Volée 5	3	7	9	8	9	4	6,67	7,5	2,58	2,11	4	6	0,39	0,53
Volée 6	3	9	9	10	9	9	8,17	9	2,56	1,72	0	7	0,31	0

Nous avons fait exprès de prendre des séries ayant la même moyenne (sauf la dernière) pour montrer la signification des autres paramètres statistiques mesurés : en bleu les paramètres centraux, en jaunes les variables de dispersion absolue, en rose les variables de dispersion relative. La plupart de ces paramètres sont calculés directement par le tableur. Il n'y a que les quartiles qu'on a dû paramétrer : pour Q3 on a tapé=QUARTILE(B2:G2;3) et pour Q1 on a tapé QUARTILE(B2:G2;1), avec cette syntaxe la médiane serait =QUARTILE(B2:G2;2), mais la fonction « Médiane » existe. Pour l'étendue, nous avons calculé =MAX(B2:G2)-MIN(B2:G2) et sinon, pour le coefficient de variation et l'écart inter-quartile relatif, il a suffit d'appliquer les définitions.



Commenter ces résultats peut paraître un peu fastidieux. C'est cependant nécessaire si l'on veut faire ressortir les aspects importants de la situation. Notre situation étant inventée, nous ne ferons pas de commentaires. Des exemples de commentaires de tableaux statistiques seront donnés plus loin. Ici, nous voulons insister sur la signification des paramètres de dispersion. Pour cela des représentations graphiques vont nous aider. La 1ère donne les résultats bruts (en ordonnée) pour les trois 1ères volées qui ont même moyenne mais des dispersions très différentes. Il peut être utile de comparer des séries triées, c'est donc cela que nous avons fait dans le 2ème graphique : les volées étant triées par résultat croissant, elles se chevauchent moins et l'on peut représenter les 6 volées sur le même graphique sans perdre de lisibilité (ou presque). Nous pouvons faire des histogrammes et regrouper les résultats par classe. Le 3ème graphique montre un tel histogramme pour les 2 premières volées, les résultats étant regroupés par classe de 2. Superposer 2 séries n'est pas forcément heureux, mais on constate ici la concentration des tirs de la volée 1 par rapport à ceux de la volée 2 qui est très dispersée. D'autres graphiques et d'autres présentation des résultats peuvent être utilisés, la palette d'expression des statisticiens étant très fournie.





Un exemple plus réaliste mais moins facile à analyser : Le tableau ci-dessous représentent la distribution des salaires par sexe dans le secteur privé et semi-public, en France sur les années 2008 à 2010. Il s'agit donc d'une répartition des données selon les déciles D1 et D9, Q1 et Q3 étant les quartiles et la médiane est indiquée par D5 (c'est aussi Q2). Pour les calculs de la moyenne ou des écarts moyens à partir de ces données déjà traitées, c'est assez difficile et aussi peu représentatif. Les centres de classes que l'on peut estimer ne sont pas les mêmes pour les hommes et pour les femmes. Un deuxième problème est qu'on ne sait pas quelle limite donner aux tranches inférieure et supérieure, on ne sait pas non plus déterminer de façon réaliste le centre de ces classes. On peut émettre des hypothèses mais cela reste des hypothèses. On ne peut pas non plus, pour la même raison calculer les étendues des deux séries. Cette représentation a le mérite de faire comprendre que les écarts sont déjà très significatifs sans les valeurs extrêmes. Elle nous montre aussi l'usage d'un indicateur de dispersion relatif : le rapport D9/D1. Il mesure la dispersion car il varie comme la dispersion et il est relatif car c'est un rapport de deux grandeurs semblables.

Distribution des salaires annuels nets de prélèvements par sexe dans le secteur privé et semi-public

	2008			2009			2010		
	Ensemble	Hommes	Femmes	Ensemble	Hommes	Femmes	Ensemble	Hommes	Femmes
D1	13 595	14 169	13 076	13 554	14 227	12 980	13 722	14 392	13 151
Q1	15 491	16 251	14 629	15 789	16 749	14 791	16 037	17 008	15 024
D5 (médiane)	**19 159**	**20 259**	**17 600**	**19 756**	**21 052**	**17 984**	**20 107**	**21 413**	**18 322**
Q3	26 136	28 262	23 401	26 869	29 135	24 000	27 345	29 607	24 462
D9	38 555	42 349	32 327	39 046	43 162	33 014	39 809	43 986	33 778
D9/D1	**2,84**	**2,99**	**2,47**	**2,88**	**3,03**	**2,54**	**2,90**	**3,06**	**2,57**

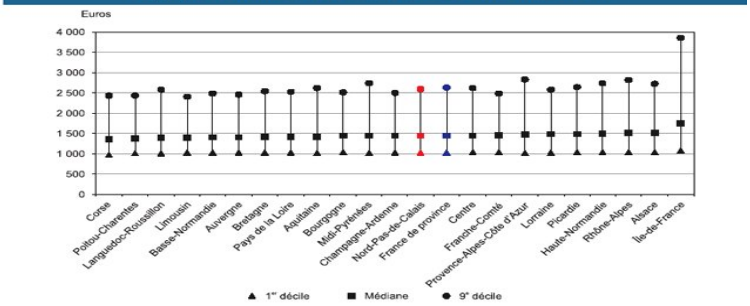
1. Y compris les chefs d'entreprise salariés.

Champ : France métropolitaine, salariés en équivalents-temps plein du secteur marchand non agricole, secteur privé et des entreprises publiques.

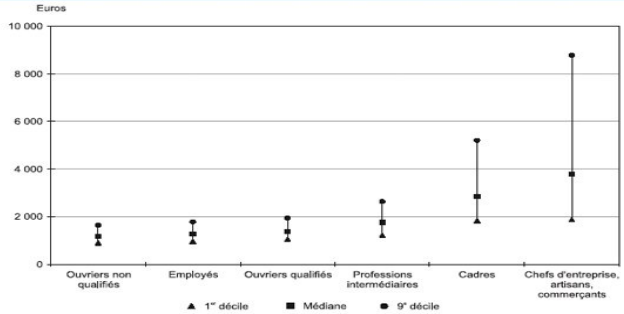
Source : Insee, DADS 2010 définitif (exploitation au 1/12).

D'une façon générale, quel usage fait-on de la dispersion ? On peut illustrer certains phénomènes économiques ou sociaux ou encore appartenant à d'autres domaines quantifiables, en voulant insister sur le plus ou moins grand étalement des données. Nous ne donnerons qu'un seul exemple : l'image ci-contre présente la dispersion des salaires en France, par région (graphique 1) et par catégorie sociale (graphique 2) en 2005. On y retrouve les déciles D1, D5 (la médiane) et D9. Les données sont exprimées d'une façon absolue (sans calculer par exemple le rapport D9/D1 qui exprimerait la dispersion relative, comme on l'a souligné plus haut). Elles sont aussi présentées de façon graphique, pour que la comparaison se fasse plus facilement : à l'œil nu (sans lecture de chiffres) les chiffres parlent d'eux-même. Le texte qui accompagne ce graphique dans la présentation qu'en a fait l'INSEE apporte quelques renseignements supplémentaires et sont aussi donnés à titre d'exemple de la prose statistique, un exercice de commentaires que l'on retrouve en dehors de l'INSEE dont c'est la vocation, dans toutes les pages économiques des journaux : *En plus de la zone d'implantation des entreprises, d'autres facteurs tant liés à l'entreprise qu'au salarié contribuent à la détermination du salaire. La catégorie socioprofessionnelle du salarié ainsi que son âge, le secteur d'activité et la taille de l'établissement influencent en très grande partie le niveau du salaire. En 2005, en Nord-Pas-de-Calais, le salaire mensuel médian des cadres s'élève à 2 846 euros soit 2,4 fois plus que celui des ouvriers non qualifiés qui s'établit à 1 178 euros. De manière plus globale, une faible proportion d'ouvriers ou employés (10%) arrivent à toucher un salaire équivalent à celui obtenu par 10% des cadres les moins rémunérés. Alors que les premiers gagnent près de 1 900 euros par mois, neuf cadres sur dix gagnent plus de 1 830 euros mensuellement. Résultante probable de l'évolution des métiers et des technologies mais aussi de la tertiarisation de l'économie, ouvriers et employés connaissent des conditions salariales proches. Ainsi, un ouvrier qualifié sur deux gagne plus de 1 383 euros nets par mois quand un employé sur deux gagne plus de 1 273 euros nets par mois.*

Graphique 1 : L'ÉVENTAIL DES SALAIRES NETS MENSUELS PAR RÉGION EN 2005



Graphique 2 : DISPERSION DES SALAIRES NETS MENSUELS PAR CATÉGORIE SOCIALE EN 2005

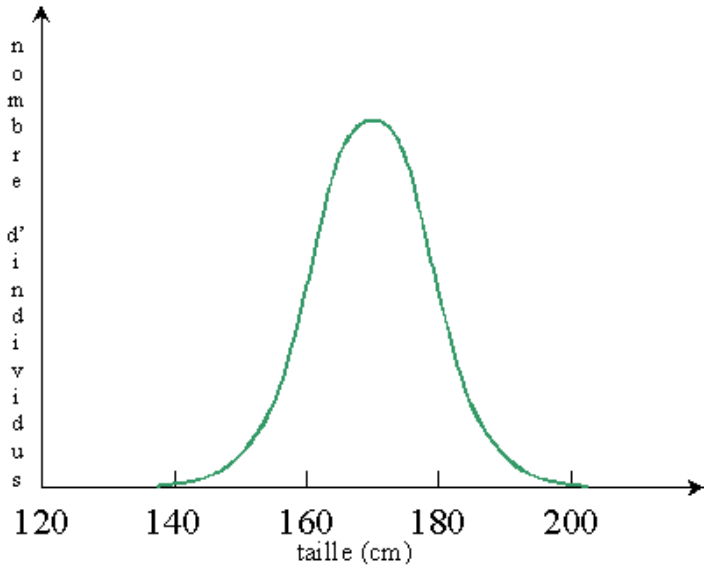


Note de lecture : le salaire mensuel moyen observé pour les commerçants, les artisans et les chefs d'entreprises de plus de 19 salariés ne reflète pas la rémunération de l'ensemble des catégories sociales mentionnées puisque les données mobilisées ne concernent que ceux ayant le statut de salariés.
Source : Insee - DADS

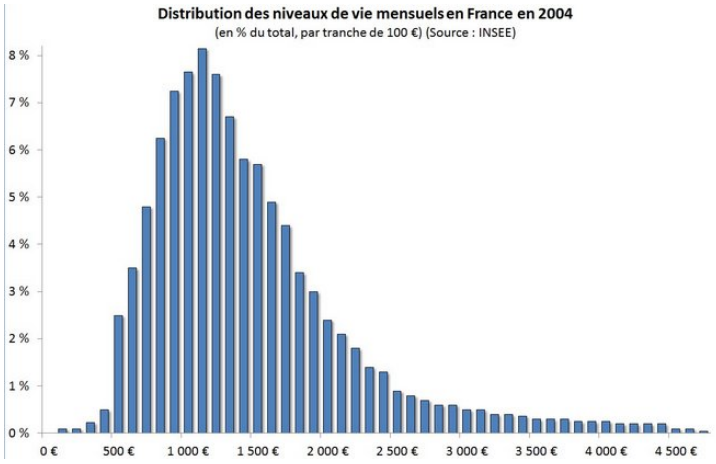
e) Asymétrie des répartitions

La plupart des répartitions sont plus ou moins symétriques, dans le sens que les valeurs se répartissent plus ou moins symétriquement par rapport aux valeurs centrales (moyenne, médiane, mode). La fameuse « courbe en cloche » qui représente cette répartition des séries symétriques est une disposition extrêmement fréquente.

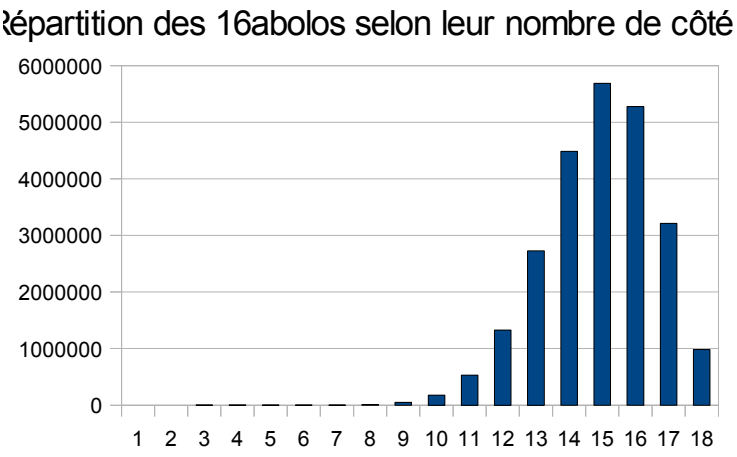
Quand on mesure la taille d'un groupe de personnes dont la taille moyenne est 170 *cm*, comme dans cet échantillon à droite, on obtient une répartition symétrique généralement autour de la moyenne. Il faut suffisamment de valeurs pour cela, mais si rien ne limite les tailles dans un sens ou dans l'autre, la répartition a toutes les chances d'être symétrique.



Si, pour une raison ou pour une autre, les valeurs peuvent s'étendre dans un sens et pas dans l'autre ; si un facteur structurel contraint les valeurs à certaines limitations ; il peut arriver qu'une distribution soit asymétrique. Ce peut être un étalement vers la droite comme dans l'illustration ci-contre où les niveaux de vie élevés se déclinent sur une large tranche alors que les faibles niveaux de vie sont restreints à une bande très étroite.

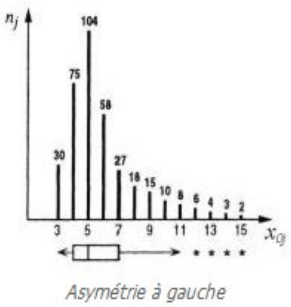


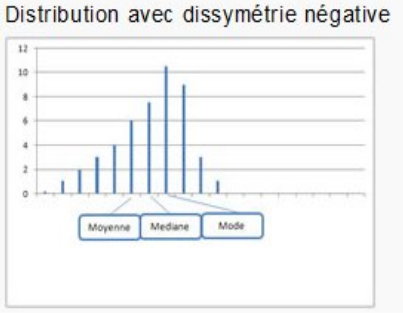
L'illustration suivante montre une répartition asymétrique dont la dissymétrie va dans l'autre sens. Les 16abolos (voir p.3) sont des formes constituées de 16 petits triangles isocèles rectangles identiques et ne peuvent, structurellement, pas dépasser 18 côtés. Les petits nombres de côtés sont possibles jusqu'à 3, mais plus rares car combinatoirement plus difficiles à générer. Il en résulte cet étalement vers la gauche.



Une conséquence de cet étalement est la séparation des paramètres centraux : ils sont confondus dans une répartition symétrique mais séparés pour les répartitions asymétriques, la moyenne étant toujours du côté de l'étalement (elle prend en compte les valeurs extrêmes) et la médiane au centre.

On peut visualiser la dissymétrie en traçant le diagramme à moustaches (boîte de Tukey) qui montre la médiane s'éloignant du quartile vers lequel il y a étalement. On peut aussi mesurer cette dissymétrie, par exemple en calculant le coefficient de Yule $Y=\frac{Q_1+Q_3-2Q_2}{Q_3-Q_1}$ qui varie entre 1 (quand $Q_1=Q_2$, étalement vers la droite maxi) et -1 (quand $Q_3=Q_2$ étalement vers la gauche maxi).

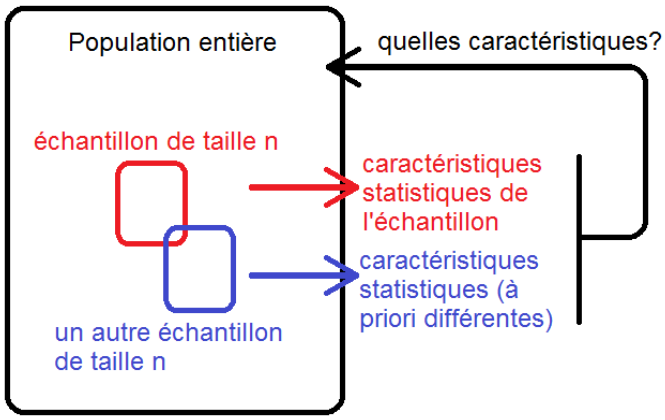




II] Échantillonnage

a) Population, échantillon et simulation

Dans certains cas, on peut connaître avec précision les caractéristiques statistiques de toute une population. Ayant accès à l'ensemble complet des individus qui composent cette population, on peut, par les méthodes de la statistique descriptive (tableaux, graphiques, paramètres tels que moyenne, médiane, etc.), analyser cette série. C'est ce que nous venons de voir dans la partie précédente. Par exemple, le recensement de la population française est un gros travail de recueil d'informations sur **tous** les individus composant la population française. Mais la population peut être trop vaste pour être étudiée dans sa totalité, ou bien les efforts qu'il faudrait consentir sont disproportionnés avec les résultats qu'on en attend. Lorsqu'on s'intéresse aux intentions de vote des Français pour une élection, on ne va pas les interroger tous (cela reviendrait à faire une élection avant l'élection). Comme il s'agit d'avoir juste une idée, on se contente d'interroger un fragment de la population : on fait un sondage sur un échantillon que l'on espère le plus représentatif de cette population (généralement 1000 à 2000 personnes suffisent). La population sur laquelle porte l'étude statistique peut même être considérée comme infinie. Dans le processus de production industrielle, on ne peut tester tous les exemplaires d'une chaîne de production. On limite donc les tests de qualité à des échantillons et l'on suppose avec certaines précautions, que les caractéristiques de l'échantillon sont valables pour la production entière. De même, si on s'intéresse aux fréquences d'obtentions de "pile" et "face" quand on joue avec le tirage d'une pièce de monnaie, le nombre de lancers de pièce à étudier est à priori infini. Il arrive aussi que la mesure d'une variable soit destructrice pour l'individu : si on étudie la durée de vie de certains appareils, il serait absurde de les faire tous fonctionner jusqu'à la panne, les rendant inutilisables. Dans tous ces cas, on est amené à n'étudier qu'une partie de la population, un échantillon, obtenu par sondage, dans le but d'extrapoler à la population entière des observations faites sur l'échantillon.



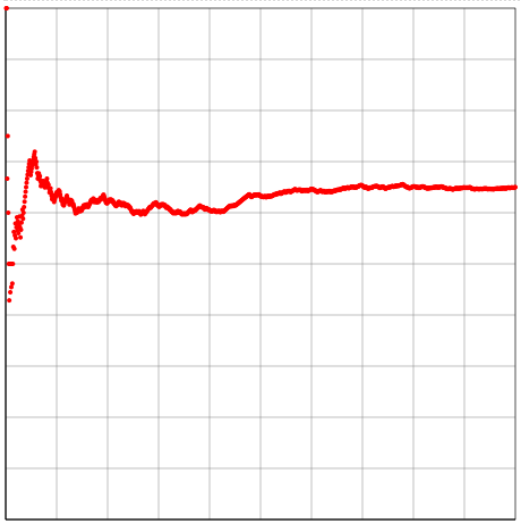
Effectuons une **expérience** pour illustrer cette question : le choix de l'échantillon entraîne t-il des fluctuations importantes des caractéristiques statistiques étudiées ? Nous effectuerons différents prélèvements dans une même population pour constituer des échantillons de même taille. Pour *simuler* cela avec un programme informatique, on peut utiliser la fonction RANDOM() qui génère un nombre aléatoire : supposons que notre population contienne 2/3 d'individus x et 1/3 d'individus non-x (x est le caractère étudié, par exemple dans un ensemble de personnes adultes, on note x la réponse « oui » à la question posée : « êtes-vous pour la continuation du programme nucléaire français? »). Lorsqu'on demande à l'ordinateur de choisir au hasard un nombre entre 1 et 100, on simule la rencontre avec un individu de cette population : si le nombre est compris entre 1 et 67 on estime qu'il est x, sinon il est non-x. On peut ainsi fabriquer un échantillon de n personnes, en recommençant n fois le tirage d'un nombre aléatoire (on prend par la suite n=1000). Notre algorithme peut se contenter de faire la somme des effectifs de la classe x pour ensuite donner la fréquence de ce caractère dans l'échantillon. Il peut aussi fournir une représentation graphique pour illustrer la formation de cet échantillon, par l'évolution graduelle des fréquences.

1. I, M, N, R et X sont des entiers.
2. X=0,
   N=1000 (nombre de tirages à effectuer),
   M=67 (taux du pourcentage correspondant à la fréquence du caractère x dans la population).
3. Pour I allant de 1 à N {
   Tirer un nombre R au hasard entre 1 et 100.
   Si (R<=M) {X=X+1}
   Placer un point de coordonnées (I;100X/I)}
4. Afficher la fréquence finale de l'échantillon : 100X/N.

L'exécution de ce programme dans Algobox conduit au graphique ci-contre et à la fréquence expérimentale de 65 (alors qu'elle était de 67 dans la population totale). C'est comme si on avait interrogé 1000 personnes et que dans cet échantillon, 65% avait le caractère *x*. Recommencer l'exécution c'est comme faire un autre prélèvement d'échantillon. Au 2ème échantillon, je trouve 66,5%. Je souhaite recommencer ainsi 200 fois le choix d'un échantillon de 1000 individus. Pour suivre ces 200 tirages, on peut modifier l'algorithme pour que, au lieu de tracer l'évolution de la fréquence dans un échantillon, il mette un point correspondant à la fréquence obtenue dans un échantillon. Le nouvel algorithme est donc :

1. I, J, M, N, R, S et X sont des entiers.
2. N=1000, M=67, S=0 (somme des fréquences dans les échantillons).
3. Pour J allant de 1 à 200 {X=0,
   Pour I allant de 1 à N {
   Tirer un nombre R entre 1 et 100.
   Si (R<=M) X=X+1}
   Placer un point de coordonnées (J;100X/N)}
   S=S+100X/N}
4. Afficher fréquence globale des 200 échantillons : S/200

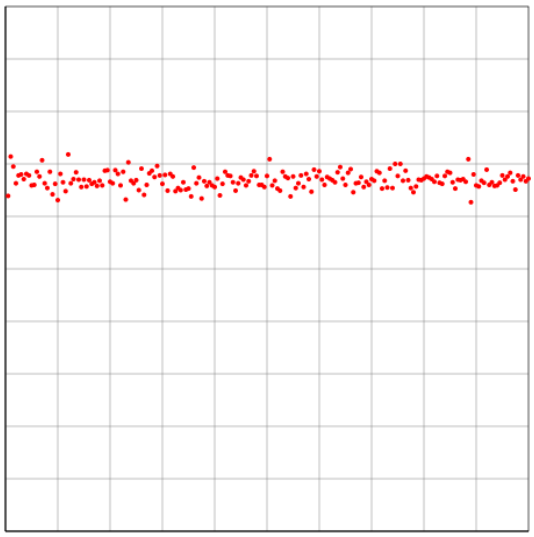
GRAPHIQUE :



Xmin: 0 ; Xmax: 1000 ; Ymin: 0 ; Ymax: 100 ; GradX: 100 ; GradY: 10

```
***Algorithme lancé***
Fréquence finale de l'échantillon : 65
***Algorithme terminé***
```

On constate alors que la fluctuation des fréquences obtenues dans chaque échantillon n'est pas très importante, sans doute à cause du choix du nombre *n* des individus dans les échantillons. Nous avions choisi une valeur importante (*n*=1000), ce qui explique sans doute cette assez grande stabilité de la fréquence. On notera tout de même que la fréquence globale obtenue auprès de ces 200 échantillons est très proche de la valeur réelle : 66,907 (au lieu de 67).

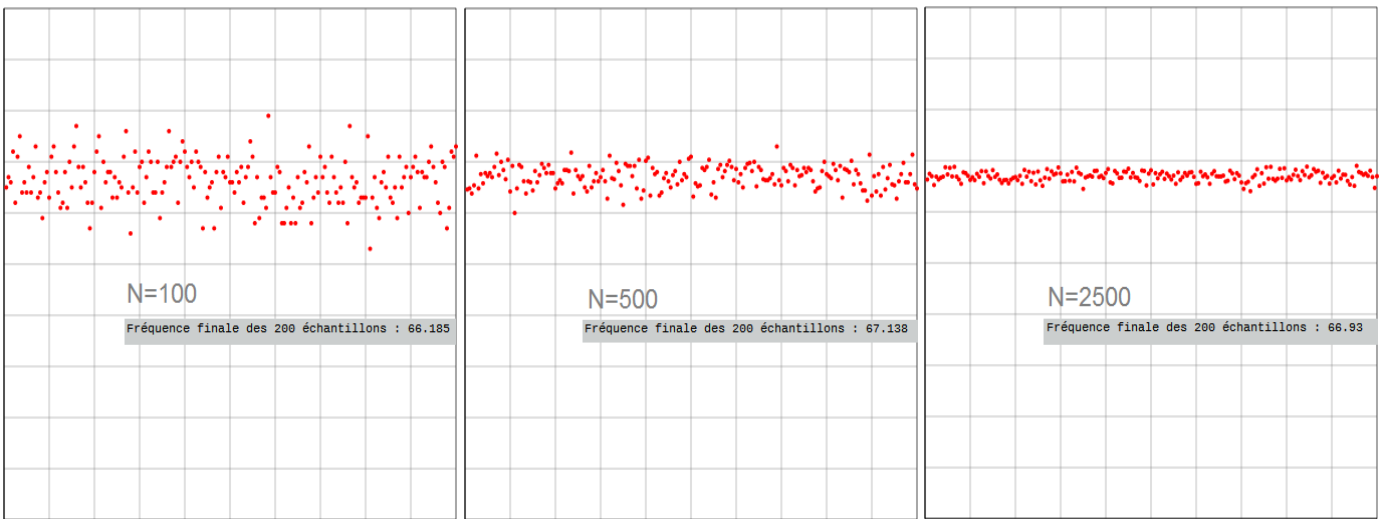


Xmin: 0 ; Xmax: 200 ; Ymin: 0 ; Ymax: 100 ; GradX: 20 ; GradY: 10

```
***Algorithme lancé***
Fréquence finale des 200 échantillons : 66.907

***Algorithme terminé***
```

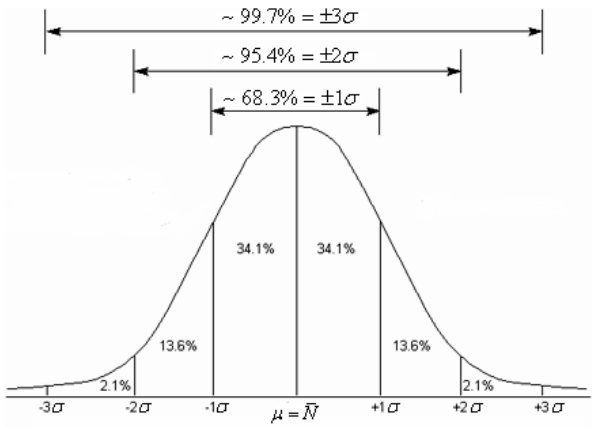
Nous pouvons maintenant tester, à l'aide du même algorithme, ce qui se passerait si on avait choisi une taille différente. Nous allons donc essayer avec *n*=100, *n*=500 et *n*=2500 pour juger de l'impact de ce paramètre (la taille *n* de l'échantillon). Bien sûr, les valeurs obtenues ici dépendent aussi du hasard : si nous lançons 2 fois de suite le même programme, il ne donnera jamais exactement la même configuration, ni le même résultat final.



Nous constatons que lorsque la taille des échantillons n augmente, les fréquences sont de plus en plus concentrées autour de la fréquence réelle, ce paramètre qui conditionne la constitution de chaque échantillon. Plus les échantillons sont grands et mieux ils représentent la population initiale. Lors de la réalisation d'un sondage, le problème est donc de déterminer la taille optimum de l'échantillon pour que l'on ne fasse pas une erreur trop grande en confondant la fréquence de l'échantillon avec la fréquence réelle de la population.

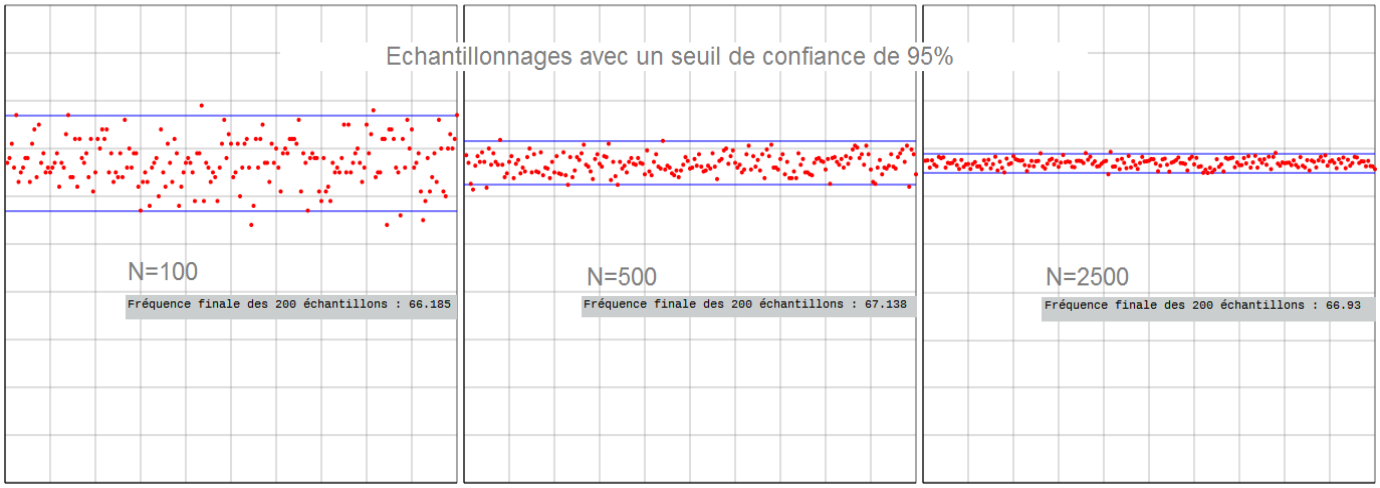
L'affirmation que l'on va extraire de l'étude de notre échantillon doit être modulée, par une formule du genre « on a 95% de chance d'avoir une fréquence réelle f comprise entre f_1 et f_2 » ou bien « on a 90% de chance d'avoir une fréquence réelle f comprise entre f_1' et f_2' » avec un intervalle $[f_1 ; f_2]$ plus restreint que l'intervalle $[f_1' ; f_2']$ car on se donne une plus petite marge d'erreur. Cette erreur que l'on assume, liée au hasard de la constitution de l'échantillon, est appelée le *taux de risque* (risque de se tromper). Si on risque de se tromper avec un taux r de 5%, alors on peut avoir *confiance* en notre estimation avec un taux c de 95%. « Se tromper » et « ne pas se tromper » sont, en effet, des évènements contraires et par conséquent on a $c=1-r$. On doit, dès lors qu'on s'est fixé un taux de risque, déterminer l'*intervalle de confiance*, c'est-a-dire $[f_1 ; f_2]$ où f_1 et f_2 sont les fréquences extrêmes entre lesquelles on doit considérer que notre fréquence réelle se trouve.

b) Intervalle de fluctuation

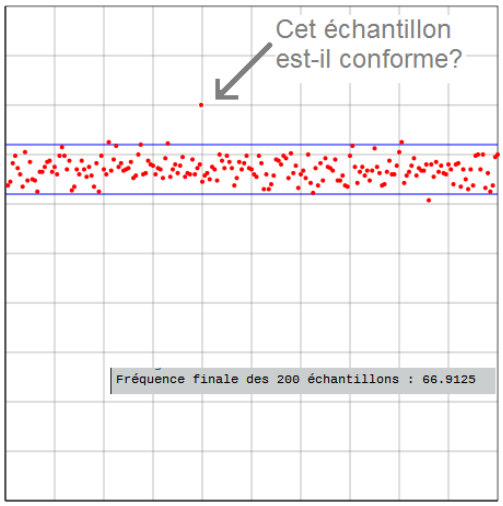


Une loi de probabilité que vous étudierez plus tard, appelée *loi normale* et connue pour sa célèbre *courbe en cloche* (aussi appelée courbe de Gauss, voir ci-contre), nous donne le résultat qui nous est nécessaire ici : Pour un risque de 5 % (valeur couramment employée), on a une très bonne approximation de l'*intervalle de fluctuation* par l'intervalle $[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}]$ où f est la fréquence réelle du caractère observé (la probabilité d'occurrence de ce caractère) et n est la taille de l'échantillon. On peut affirmer que la fréquence expérimentale f_e se situe dans cet intervalle avec un risque très faible (dans 5 cas sur 100) de se tromper. Ces résultats s'appliquent aux cas où la fréquence n'est ni trop grande, ni trop petite (f doit être comprise entre 0,2 et 0,8) et pour des échantillons pas trop petits ($n\geq 25$).
Par exemple, avec n=100, on a l'intervalle de fluctuation $[f-0{,}1;f+0{,}1]$, avec n=500, on a l'intervalle de fluctuation $[f-0{,}045;f+0{,}045]$ approximativement et pour n=2500, on a l'intervalle de fluctuation $[f-0{,}02;f+0{,}02]$.

Traçons pour illustrer cela, sur nos précédents schémas, les droites horizontales représentant les limites dans lesquelles devraient se situer les fréquences expérimentales avec un seuil de confiance de 95% pour n=100, n=500 et n=2500.



Exemple d'application : On sait que dans une population donnée, il y a 67% de fumeurs, soit une fréquence f=0,67 pour le caractère x=« être fumeur ». Sur 400 malades atteints d'un cancer des bronches, on trouve 320 fumeurs, soit une fréquence de 0,8. Un tel résultat permet-il d'affirmer que le tabagisme augmente les chances d'avoir un cancer des bronches ? Il faudrait savoir si la différence entre 0,8 et 0,67 est significative vu la taille de l'échantillon prélevé dans la population. Si fumer ne donne pas le cancer, on devrait avoir une fréquence dans l'échantillon, contenue à l'intérieur de l'intervalle de fluctuation. Or, pour un échantillon de 400 personnes et une fréquence f de 0,67 l'intervalle de fluctuation est $[f-0{,}05\,;f+0{,}05]$, soit $[0{,}62\,;0{,}72]$.



La fréquence du cancer au sein de cet échantillon de la population, n'est pas conforme à ce qu'on est en droit d'attendre d'un échantillon pris au hasard ($0{,}8\notin[0{,}62\,;0{,}72]$). Par conséquent, on est en droit d'affirmer, avec une probabilité de se tromper très faible (moins de 5%), que fumer augmente le risque d'attraper un cancer des bronches. Notre illustration propose une mise en scène : sur un tirage de 200 échantillons de 400 personnes présentant une probabilité de 67% d'être fumeur, la fréquence expérimentale du nombre de fumeur se répartie entre 62% et 72% dans 95% des cas. Quelques rares cas sortent de cet intervalle de fluctuation, on en voit quelques-uns, mais ils ne s'écartent pas beaucoup de l'intervalle. Notre échantillon de malade est très loin de l'intervalle : nous l'avons ajouté manuellement, il n'est pas sorti au hasard (c'est un trucage!). Si la loi normale n'exclut pas, à priori, qu'un échantillon pris au hasard dans une population soit très différent de ce que l'on peut attendre (on parle de valeur aberrante), elle en calcule la probabilité d'occurrence. Ici cette probabilité serait certainement inférieur à 1% (quasi-impossible).

c) Intervalle de confiance

Lorsqu'on effectue un sondage, on ne connait pas à priori, la fréquence réelle f. On mesure une fréquence dans l'échantillon, notons-là f_e, et l'on voudrait estimer un intervalle de confiance au seuil de 95%. Dans ce cas, le résultat énoncé s'applique avec f_e à la place de f. On ainsi un *intervalle de confiance* qui est $[f_e-\frac{1}{\sqrt{n}}\,;f_e+\frac{1}{\sqrt{n}}]$ et à l'intérieur duquel on peut s'attendre à trouver la fréquence réelle f avec la probabilité de 95%. La raison en est simple : si la loi prévoit que $f_e\in[f-\frac{1}{\sqrt{n}}\,;f+\frac{1}{\sqrt{n}}]$, c'est que $f-\frac{1}{\sqrt{n}}<f_e<f+\frac{1}{\sqrt{n}}$, mais en ajoutant $\frac{1}{\sqrt{n}}$ à l'inégalité de gauche on obtient $f<f_e+\frac{1}{\sqrt{n}}$ et en retranchant $\frac{1}{\sqrt{n}}$ à l'inégalité de droite on obtient $f_e-\frac{1}{\sqrt{n}}<f$ ce qui conduit bien à l'encadrement $f_e-\frac{1}{\sqrt{n}}<f<f_e+\frac{1}{\sqrt{n}}$.

Exemple d'application : Une enquête préliminaire à l'implantation d'un supermarché dans une petite ville a montré que, sur 900 familles interrogées au hasard, 270 envisageaient d'y aller en voiture le samedi après-midi (jour de principale affluence). Sachant que la population de cette ville compte 2000 familles, quel est le nombre minimal de place de parking à prévoir pour que, dans 95% des cas, il ne soit pas complet le samedi après midi ? La fréquence f_e dans l'échantillon est $f_e=\frac{270}{900}=0{,}3$. L'intervalle de confiance au seuil de 95% a donc du sens ($0{,}3\in[0{,}2\,;0{,}8]$ pour la fréquence et $n>25$ pour l'effectif) et il est $[0{,}3-\frac{1}{\sqrt{900}}\,;0{,}3+\frac{1}{\sqrt{900}}]$ soit, $[0{,}267\,;0{,}333]$ approximativement. Dans 95% des cas, il y aura donc moins de 33,3% (pourcentage correspondant à 0,333) de la population totale qui ira au supermarché, soit moins de 667 familles. Il faudrait donc, si l'on en croit cette étude, prévoir 667 places de parking pour être *sûr à 95%* qu'il ne sera pas complet.

Utilisation d'un tableur : Nous avons fait toutes nos simulations sur Algobox, mais le tableur permet aussi bien d'en réaliser. On peut recopier la formule écrite dans une cellule, par exemple le tirage d'un nombre aléatoire avec la fonction ALEA() qui est l'équivalent de RANDOM() sur le tableur *calc* d'OO, afin qu'elle s'applique n fois (taille de notre échantillon). L'ensemble des tirages est modifiable en utilisant la combinaison de touche F9 (sous Windows) ce qui permet de choisir un nouvel échantillon.
Exemple : On veut simuler le lancer d'une pièce de monnaie. On a besoin de générer un nombre aléatoire

égal à 0 (pour « pile ») ou à 1 (pour « face »). Supposons que l'on veut réaliser 50 tirages d'une pièce, on va recopier 50 fois notre formule. Quel est le pourcentage de « pile » obtenu lorsqu'on tire 50 fois la pièce ? On les compte tout simplement, avec la fonction NB.SI(*plage*,*valeur*) où « plage » est l'ensemble des cellules concernées par le comptage, et « valeur » est la valeur dont on comptabilise les effectifs. On obtient ainsi, très facilement, les fréquences de l'évènement PILE dans notre échantillon de 50 tirages.

C3 =NB.SI(B$2:B3;"PILE")

	A	B	C	D
1	N°individu	Valeur	Total PILE	Fréquence PILE dans l'échantillon
2	1	PILE	1	0,5200
3	2	PILE	2	
4	3	FACE	2	

B3 =SI(ALEA.ENTRE.BORNES(0;1)=0;"PILE";"FACE")

	A	B	C	D	E	F
1	N°individu	Valeur	Total PILE	Fréquence PILE dans l'échantillon		
2	1	PILE	1	0,5800		
3	2	PILE	2			
4	3	FACE	2			

Pour simuler avec un tableur le tirage de 200 échantillons de 50 lancers d'une pièce, comment fait-on ? On peut faire 200 fois le retirage avec la touche F9 (pour la fonction ALEA.ENTRE.BORNES(*min,max*) il faut utiliser la combinaison Ctrl+Maj+F9) et noter à chaque fois le résultat obtenu dans une cellule, mais cela est bien fastidieux! On trouvera sans doute plus intéressant la possibilité de recopier la formule obtenue sur une colonne, sur les 199 colonnes qui se trouvent à droite. De cette façon, on pourra tout retirer d'un seul coup avec la touche F9 ou la combinaison Ctrl+Maj+F9 !

N°individu	Valeur E1	Valeur E2	Valeur E3	Valeur E4
1	FACE	FACE	FACE	FACE
2	FACE	PILE	PILE	FACE
3	FACE	PILE	FACE	PILE

=SI(ALEA.ENTRE.BORNES(0;1)=0;"PILE";"FACE"

48	PILE	PILE	PILE	
49	FACE	FACE	PILE	
50	FACE	FACE	FACE	=NB.SI(H2:H51;"PILE")
Total 50 E.	1214	23	25	26
fréq.moy.	48,56			

Remarques sur les fonctions RANDOM() et ALEA() : ces fonctions génèrent un nombre pseudo-aléatoire compris entre 0 et 1. En réalité, 1 n'est jamais fourni par cette fonction qui génère des nombres de l'intervalle [0;1[, c'est-a-dire que si on note x le nombre généré par RANDOM() alors on a $0 \leq x < 1$. On pourra obtenir 0,999000 ou 0,999999 mais jamais 1. Donc si on veut obtenir un nombre entre 0 et 100 et que l'on multiplie RANDOM() par 100 on obtient un nombre $0 \leq 100x < 100$ qui peut être égal à 99,9 ou 99,9999 mais pas à 100. Pour ramener ce nombre décimal à un entier, il nous faut tronquer le résultat. La troncature est cette opération qui consiste à supprimer la partie décimale d'un nombre. Elle se note souvent avec la fonction ENT() ou FLOOR(). Ainsi ENT(99,9)=99. Si on définit notre nombre aléatoire par la combinaison ENT(100*RANDOM()), on va se retrouver avec des nombres entiers aléatoires compris entre 0 et 99. Pour obtenir ce que l'on souhaitait dans notre exemple (un nombre aléatoire compris entre 1 et 100), il suffit donc d'ajouter 1 au résultat. Ainsi ENT(100*RANDOM())+1 est l'instruction correcte qui permet d'obtenir ce que l'on est en droit d'attendre. Sur Algobox on tapera plutôt : floor(100*random())+1.
De la même façon, pour obtenir une simulation de PILE ou FACE, il faut donc taper l'instruction ENT(2*RANDOM()). On obtiendra ainsi les nombres 0 ou 1 d'une façon aléatoire, ce que l'on transforme en PILE (si le nombre est 0) ou FACE (si le nombre est 1). L'instruction ENT(2*RANDOM())+1 donnerait des nombres égaux à 1 ou 2. Attention aussi à ne pas confondre ENT() la partie entière d'un nombre et ROUND() l'arrondi. Avec une instruction comme ROUND(2*RANDOM()), on obtiendrait des nombres égaux à 0 (dans 25% des cas), 1 (dans 50% des cas) et 2 (dans 25% des cas).

Cela est un peu compliqué ? En effet, surtout que l'on a souvent besoin de générer des nombres aléatoires entiers. Les deux faces d'une pièce : 0 ou 1, les 6 résultats d'un dé cubique : 1, 2, 3, 4, 5 ou 6, les 12 mois de l'année, les 7 petits nains, etc. Pour cette raison, la plupart des langages de programmation, celui du tableur et Algobox compris, offre une autre fonction qui s'appelle, sur le tableur *calc* d'OO ALEA.ENTRE.BORNES(*min;max*) ou « *min* » et « *max* » sont des valeurs incluses dans les résultats possibles. Sur Algobox, cette fonction est ALGOBOX_ALEA_ENT(*min,max*). Le tirage d'une pièce sera simulé par l'instruction ALEA.ENTRE.BORNES(0;1) et celui d'un dé par ALEA.ENTRE.BORNES(1;6). C'est alors beaucoup plus facile n'est-ce-pas ?