INTRODUCTION A LA STATISTIQUE DESCRIPTIVE

L1ECO

Septembre 2014

”if you want to have a long life, get a wife”-(vérité statistique)

1 INTRODUCTION

”Dans ce pays, une loi fixe à l’avance combien de personnes se marieront et à quel âge, combien de jeunes filles épouseront des hommes âgés, combien de jeunes hommes des femmes âgées, pour combien de couples la différence d’âge sera de tel ordre, pour combien elle sera de tel autre ordre, combien de veufs épouseront des veuves, combien de divorces seront prononcés par les tribunaux, etc....

Ce qui ne pourrait jamais être réalisé de cette manière par la volonté et par la force des hommes, s’accomplit merveilleusement, sans aucune intervention, grâce à l’organisation naturelle de la société humaine.....

En examinant les mariages, les suicides, les crimes et en dégageant leurs lois, nous pouvons prévoir avec une très grande exactitude combien de mariages, de divorces, de suicides et de crimes se produiront pendant une année et comment ils se répartiront.

L’examen futur des résultats de cette année révélera qu’ils sont tout aussi conformes aux prévisions que si nous nous trouvions dans ce Etat imaginaire. Le plus étonnant est que nous soyons nous-mêmes les éléments d’un grand mécanisme et que nous gardions l’entière liberté de nos mouvements sans pour autant empêcher le fonctionnement préétabli de ce mécanisme.” Adolph Wagner 1864

A l’origine, la statistique consistait en une simple collecte de chiffres, ce qui correspond à la signification première de state-istique, ensemble d’informations concernant la population et l’économie, indispensable à l’Etat. La tâche initiale de la statistique a été, comme le dit Cournot : ” le recueil des faits auxquels donne lieu l’agglomération des hommes en sociétés politiques ”.

Actuellement, la statistique s’est développée comme une méthode scientifique d’analyse s’appliquant à l’économie et à l’ensemble des sciences sociales et de la nature.

L’assurance vie est la première grande application du calcul des probabilités. Avant la fin du 17ème siècle, Halley établit les premières tables de mortalité, qui lui permirent d’estimer la durée de vie. Se prémunir contre l’adversité a un prix depuis deux millénaires, ce prix étant calculé auparavant sans règle précise, en s’appuyant sur l’expérience. En 1637, avant la naissance de Halley, un savant crétois, du nom de Canopius, se prépare un breuvage que l’on goûte pour la première fois en Angleterre : un café..Le succès est tel que s’ouvrent bientôt des centaines de ”coffee house”, dont l’un donnera naissance à la célèbre Lloyd’s. Halley, quant à lui, utilisa l’outil de prédilection des assureurs, la loi des grands nombres, loi indispensable à l’exercice de l’activité d’assurance. Cette loi fait le lien entre les statistiques et les probabilités, puisqu’elle met en relation fréquence statistique et probabilité. On devra s’y résoudre ou s’en réjouir : le hasard est une notion qui ne s’analyse pas sans mathématiques...et un contrat d’assurance consiste à fixer le prix d’une prestation, sorte ”d’antimarchandise”que constitue le risque, le montant de la prime consistant en une évaluation du hasard.

On notera au passage, qu’un contrat d’assurance est d’une nature très particulière, son exécution faisant toujours apparaître des conflits d’intérêts a postériori : si la cargaison d’un bâteau arrive à bon port, son affréteur trouve la prime d’assurance trop coûteuse ; si un naufrage se produit, l’assureur trouve la prime insuffisante.

Si du point de vue mathématiques, l’assurance est assimilable à un jeu de hasard, il est clair que l’assureur n’est en aucun cas un joueur. Selon le principe indemnitaire, l’assurance ne peut permettre à l’assuré de s’enrichir, mais seulement de ”compenser” une perte. Par ailleurs, le risque doit être assurable, et celà demande que plusieurs conditions soient remplies :

- L’assureur doit être en mesure de déterminer la probabilité de réalisation du risque.
- Les risques assurés doivent être suffisamment nombreux afin que la mutualisation puisse fonctionner et que **la loi des grands nombres** puisse s’appliquer.
- L’assureur doit veiller à ce qu’il y ait une bonne dispersion des risques, afin d’éviter des points d’accumulation, c’est-à dire des sinistres multiples ayant la même origine ; les risques doivent être suffisamment éparpillés.
- Enfin les risques doivent être **indépendants** entre eux.

L’analyse statistique étudie un ensemble (**ou population**) d’unités (**éléments, individus**) envisagés sous l’angle d’un

caractère donné.

2 Statistique descriptive et statistique inférentielle

Le présent cours aborde deux aspects de la statistique, d'une part la statistique descriptive, en ce sens que l'on va étudier les graphiques,les paramètres, les outils permettant de présenter et d'analyser l'information de façon synthétique et utilisable et d'autre part la statistique inférentielle qui a pour but de généraliser l'information basée sur des échantillons à la population complète, ou selon l'expression consacrée de faire des inférences.

Il est clair qu'en général on n'a pas accès à toute la population et que l'on doit travailler sur un échantillon. Si l'on veut savoir si la soupe est correctement salée, on ne mangera pas toute soupe contenue dans la soupière, mais on en goûtera une cuillère. La statistique inférentielle estime par exemple le pourcentage d'électeurs prêts à voter pour tel candidat, à partir de l'étude d'un échantillon de 1000 électeurs.

Evidemment si la soupe est mal mélangée, cet échantillon ne sera pas représentatif de la population (échantillon biaisé) et ne donnera pas une idée exacte quant à la salaison de la soupe. Le prélèvement d'échantillons, que nous ne traiterons pas ici, doit être effectué suivant des méthodes rigoureuses assurant l'obtention d'un échantillon représentatif. Si l'on s'intéresse aux notes d'un partiel dans un EC donné, on ne peut prendre comme échantillon les étudiants du premier rang, ni ceux du dernier rang, ni ceux dont le nom commence par un A, tous ces échantillons sont biaisés, contrairement au mode le d'échantillonnage le plus important, le mode aléatoire simple, qui consiste à effectuer des tirages aléatoires et indépendants.

En plus du mode d'échantillonnage, il est important de s'intéresser à un élément important : la taille de l'échantillon. On retiendra le résultat, à priori surprenant : c'est la taille de l'échantillon qui conditionne la qualité des résultats, et ce sans rapport avec la taille de la population totale. Avec un échantillon représentatif de 100 étudiants parmi les 20000 étudiants d'une université, on représente la population étudiante de l'université avec une qualité équivalente à celle obtenue par un échantillon représentatif de 100 votants parmi les 131.2 millions de votants à l'élection présidentielle américaine de 2008.

3 Dictionnaire de base

1. Une **population** est l'ensemble des éléments auquels se rapportent les données étudiées (étudiants d'une université, habitants d'un pays, entreprises d'un secteur...).
2. Dans une population donnée, chaque élément est appelé un "**individu** " ou une "unité statistique".
3. En fait, la collecte d'informations sur une population est rarement effectuée de façon exhaustive (enquête sur la totalité des individus) ; on a souvent recours à des enquêtes par sondage qui portent sur une partie de la population, appelée **échantillon**.
4. Il existe deux types de caractères statistiques : les caractères **quantitatifs**, c'est à dire qui prennent des valeurs numériques (taille, salaire, etc...) et les caractères **qualitatifs** (sexe, métier, couleur des yeux, situation matimoniale, etc..), ceux dont les modalités ne sont pas numériques.
5. **Caractère quantitatif discret** : qui prend des valeurs isolées. Exemple : nombre d'enfants; les valeurs sont des entiers naturels.
6. **Caractère quantitatif continu :** prend des valeurs quelconques dans un intervalle ; les données sont regroupées en classes ;
   exemple : salaires, taille ...
7. **Modalités** : les modalités d'un caractère sont ses différentes "valeurs" ; exemple : l'état matrimonial comporte souvent cinq modalités : célibataire, marié, pacsé, veuf, divorcé.
8. **Effectif :** l'effectif d'une modalité, en général noté n_i, représente le nombre d'individus correspondant à cette modalité. En général, on note $n = \sum n_i$, l'effectif total de la^population ou de l'échantillon sur lequel on travaille.$(\sum :$ lu "Sigma" et signifie somme$)$.

9. **Effectifs cumulés croissants : notation :** $n_i cc$. On effectue la somme des effectifs des modalités inférieures ou égale à une modalité donnée ; exemple des notes à un devoir : si l'on suppose que les notes sont des entiers de 1 à 10, l'effectif cumulé croissant correspondant à 8, consiste à compter le nombre de personnes ayant une note inférieure ou égale à 8.(cf exemple)

10. **Effectifs cumulés décroissants** : **notation :** $n_i cd$. On effectue la somme des effectifs des modalités supérieures ou égales à une modalité donnée; pour le même exemple que précédemment, on trouvera pour 8, le nombre total de personnes dont la note est supérieure ou égale à 8.

11. **Fréquence (relative) :** la fréquence d'une modalité, notée f_i, est donnée par: $f_i = \dfrac{n_i}{N}$; elle représente la proportion d'individus se rapportant à une modalité par rapport à l'effectif total. On a :
$\boxed{0 \leq f_i \leq 1 \text{ et } \sum f_i = 1}$; enfin une fréquence peut être donnée en pourcentage et alors, la somme des fréquences donne 100%.

12. **Fréquences cumulées croissantes ou décroissantes : notations :** $f_i cc$ et $f_i cd$
$f_i cc = \dfrac{n_i cc}{n}$ et $f_i cd = \dfrac{n_i cd}{n}$.

13. **Amplitude de classe** : l'amplitude de la classe $[a\,;b[$ est $b-a$, c'est la longueur de l'intervalle.

14. **Centre de classe :** C'est le milieu de l'intervalle, donc le centre de $[a\,;b[$, soit,la moyenne arithmétique x de a et b : $x = \dfrac{a+b}{2}$.

15. **Densité de classe** (continu) : penser à la densité en démographie.
    a. On appelle densité d'effectif de la classe $[a_i\,;b_i[$, le nombre d_i défini par : $d_i = \dfrac{n_i}{b_i - a_i}$, qui représente le nombre d'individus par unité de classe. Si par exemple, lors d'un concours on a 40 candidats dont la moyenne est dans la classe $[10, 15[$,la densité de cette classe est de $\dfrac{40}{5}$,soit de 8 ; celà signifie qu'en moyenne, dans cet intervalle, il y a 8 étudiants par point ; l'hypothèse étant la répartition uniforme à l'intérieur d'une classe, on a en théorie, 8 étudiants entre 10 et $11, 8$ entre 11et 12,etc, 8 entre 14 et 15.

    b. On appelle densité d e fréquence de la classe $[a_i\,;b_i[$, le nombre d'_i défini par : $d'_i = \dfrac{f_i}{b_i - a_i}$, qui représente le pourcentage d'individus par unité de classe. S i dans l'exmple précédent il y a en tout 500 candidats, alors la fréqence de la classe $[10, 15[$est $\dfrac{40}{500} = 0.08$, et la densité de fréquence $d'_i = \dfrac{0.08}{5} = 0.016$,qui signifie que dans cette classe, il y a en moyenne 1.6% étudiants par point.

16. **Mode**
    a. Caractère discret : un mode est une valeur du caractère ayant l'effectif maximal ; on notera qu'une série statistique peut avoir un mode (unimodale) ou des modes (plurimodale). Dans l'exercice 1 le mode est 2.

    b. Caractère continu : une classe modale est une classe ayant la densité maximale. Dans l'exemple 2, il s'agit de la classe $[110\,;120[$ et dans l'exercice 3 de la classe $[10;\ 12[$.

17. **Médiane**

    La médiane est à priori la valeur du caractère qui partage la série ordonnée (de la plus petite modalité à la plus grande) en deux groupes de même effectif ; on doit cependant distinguer plusieurs situations. On utilisera les effectifs ou fréquences cumulées croissantes.

18. **Quartiles**

    Il y a trois quartiles : Q_1, Q_2,et Q_3 ; le principe est le même que pour la médiane, mais il s'agit de partager la série en quatre groupes comprenant 25% de la population. Q_1 est la plus petite valeur telle qu'il y ait au moins 25% des valeurs de la série inférieures ou égales à Q_1. Q_2 est la médiane et Q_3 est la plus petite valeur telle qu'il y ait au moins 75% des valeurs de la série inférieures ou égales à Q_3. On distingue comme pour la médiane le cas discret et le cas continu et on utilise les effectifs ou fréquences cumulées croissantes.

    Boîte à moustaches ou Box plot : ce graphique sur lequel nous reviendrons utilise les quartiles et est très précieux pour comparer diverses séries (salaires dans différents pays européens par exemple).

19. **Moyenne**

La moyenne d'un caractère statistique quantitatif x est notée $\overline{x}$ et définie par :

$$\boxed{\overline{x} = \frac{1}{n}\sum n_i x_i = \sum f_i x_i}$$

On notera que : $\sum n_i x_i = n\overline{x}$.Par exemple si une entreprise comporte 300 salariés et que le salaire moyen est de 1500 euros, la masse salariale est : $\sum n_i x_i = n\overline{x} = 300 * 1500 = 450000$ euros.

Cas particulier : la moyenne arithmétique simple de n observations x_i est donnée par : $\boxed{\overline{x} = \frac{1}{n}\sum x_i}$ (les effectifs n_i sont égaux à 1).

20. **Variance et écart-type**

Pour mesurer les fluctuations d'un caractère autour de sa moyenne, c'est-à dire fournir un indicateur de dispersion, on définit la variance et l'écart-type, définis respectivement par :

$$\boxed{V(x) = \frac{1}{n}\sum n_i (x_i - \overline{x})^2 = \frac{1}{n}\sum n_i x_i^2 - \overline{x}^2}$$

($MC - CM$: moyenne des carrés moins carré de la moyenne ; on note que $V(x) \geq 0$).

$\sigma\,(x) = \sqrt{V(x)}$.

4 LES PRINCIPAUX GRAPHIQUES

4.1 Caractères qualitatifs ou quantitatifs discrets

1. Le diagramme en bâtons ou tuyaux d'orgue.

   -On peut "épaissir" les bâtons, jusqu'à en faire des rectangles, mais ils ne doivent pas être "collés", comme dans les histogrammes (caractères continus).

   -Préciser les axes et la nature du diagramme : effectifs, fréquences...

2. Le diagramme circulaire (camembert)

   Il doit comporter une légende, être colorié et le calcul des angles doit figurer sur la copie : 360° correspondant à une fréquence de 100%, l'angle correspondant à la modalité x_i,de fréquence f_i,sera en degrés : $\alpha_i = 360 f_i$ ($180 f_i$ dans le cas d'un "demi camembert") ; ainsi une modalité de fréquence 50%,sera représentée par un demi disque ($180°$).

4.2 Caractères quantitatifs continus

1. **L'histogramme**

   Les données sont regroupées en classe et le diagramme utilisé couramment est l'histogramme **(**des effectifs ou des fréquences**).**

   Chaque classe (portée en abscisse) est représentée par un rectangle basé sur cette classe, les rectangles étant donc "collés" car on est en présence

   d'un caractère continu.

   Le principe : les effectifs (ou les fréquences) sont représentés par **l'aire** de ces rectangles ; les hauteurs de ces rectangles ne sont donc pas nécessairement égales aux effectifs ; plus précisemment, dans le cas où les amplitudes de classes sont inégales (donc les bases des rectangles inégales), les hauteurs doivent être calculées de façon à ce que les aires des rectangles soient proportionnelles aux effectifs. Regardons les deux cas :

   a. **Classes de même amplitude** : la simplicité

      Rappel : pour une classe de type $[a\,;b[$, l'amplitude est la longueur de l'intervalle, c'est à dire, $b - a$. Exemple : $[75\,;99[$ a pour amplitude 24, soit $99 - 75$.

      Dans le cas où les amplitudes de classe sont égales, on prend pour hauteurs des rectangles les effectifs, les aires des rectangles étant proportionnelles aux hauteurs.

Exemple à traiter : statistiques du nombre d'articles vendus dans une journée selon leur prix.

Classes	Effectifs n_i
$[0; 200[$	1500
$[200; 400[$	1000
$[400; 600[$	800
$[600; 800[$	500

Représenter l'histogramme des effectifs.

b. **Classes d' amplitudes différentes**

On va devoir choisir une unité d'amplitude de classes et déterminer les hauteurs des rectangles pour respecter le principe énoncé plus haut : ce sont les aires des rectangles qui doivent être proportionnelles aux effectifs (ou aux fréquences). Les hauteurs des rectangles seront appelées **les effectifs corrigés** ; ces effectifs corrigés sont proportionnels aux **densités** de chaque classe, c'est à dire aux effectifs par unité d'amplitude. On

doit avoir présent à l'esprit l'exemple de la démographie, la densité représentant le nombre d'habitant au km^2.

On définit donc **la densité** (d'effectif ou de fréquence) d'une classe par :

$d_i\,(n_i) = \dfrac{n_i}{A_i}$, où A_i désigne l'amplitude de la classe $[a_i\ ;\ b_i[$, avec $A_i = b_i - a_i$

ou $d_i\prime = d_i\,(f_i) = \dfrac{f_i}{A_i}$pour la densité de fréquence.

Une convention courante est de prendre pour hauteurs des rectangles (effectifs corrigés ou fréquences corrigées) le produit des densités par l'amplitude minimale de classe .

On notera, $n_i cor$, les effectifs corrigés et $f_i cor$,les fréqneces corrigées.

Dans l'exemple qui suit, l'amplitude minimale de classe est 100 ; on a multiplié les densités par 100 pour obtenir les effectifs corrigés.

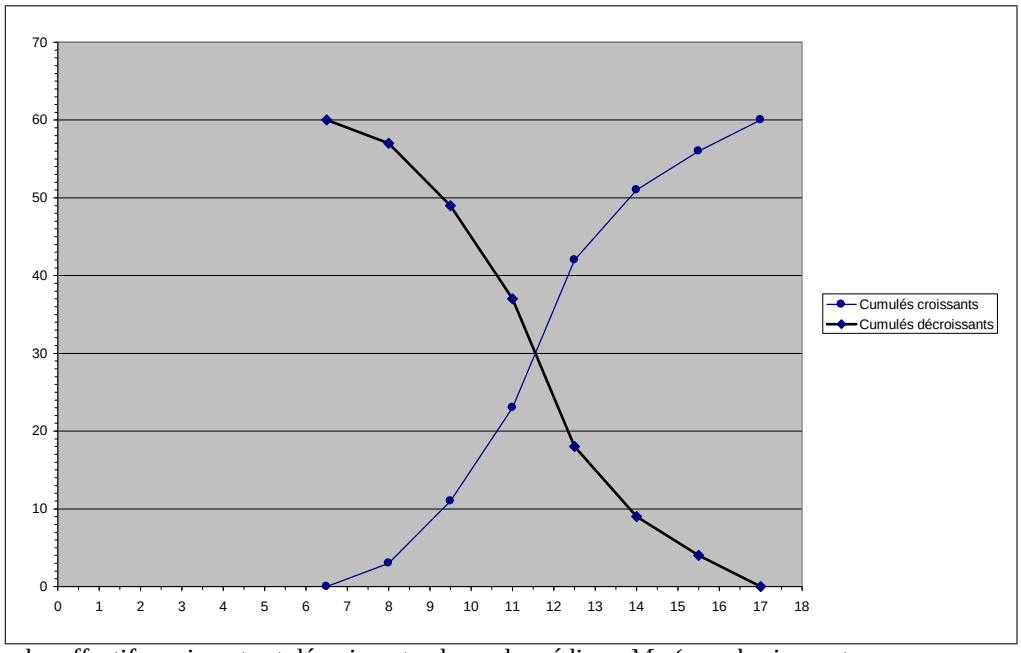
Classes	a_i	b_i	n_i	A_i	d_i	$n_i cor$	f_i	d'_i	$f_i cor$
$[0; 100[$	0	100	1500	100	15	1500	0.394 7	0.003 947	0.394 7
$[100; 500[$	100	500	1000	400	2.5	250	0.2632	0.000 658	0.0658
$[500; 1000[$	500	1000	800	500	1.6	160	0.210 5	0.000 421	0.0421
$[1000; 1500[$	1000	1500	500	500	1	100	0.131 6	0.000 263 2	0.0263
			$n = 3800$				$\sum f_i = 1$		

Interprétation des densités (par exemple sur la premère classe) : sous l'hypothèse de répartition uniforme des individus statistiques à l'intérieur d'une classe, dans la classe $[0; 100[$,il y a en moyenne 15 individus par unité de classe, ce qui représente $0.3947\ \%$ de la population totale.

Tracer l'histogramme des effectifs. On notera que celà revient à utiliser une unité d'amplitude de classe, 100, et à mesurer les classes avec cette amplitude. La première classe dont la base vaut ainsi 1 (mesurée avec notre unité) a une hauteur qui vaut son effectif (1500), puisque l'aire du rectangle est : $1 * 1500 = 1500$, le deuxième rectangle a une base de 4 donc sa hauteur sera divisée par 4, son aire étant : $4 \times \dfrac{1000}{4} = 1000$ (l'effectif), etc.

2. **Le Polygone des effectifs cumulés croissants** (ou décroissants)

Classes	n_i	n_icc	n_icd
$[6.5; 8[$	3	3	60
$[8; 9.5[$	8	11	57
$[9.5; 11[$	12	23	49
$[11; 12.5[$	19	42	37
$[12.5; 14[$	9	51	18
$[14; 15.5[$	5	56	9
$[15.5; 17[$	4	60	4
Total	60		



Le point M, intersection des polygones des effectifs croissants et décroissants, donne la médiane Me (son abscisse est Me et son ordonnée $\frac{n}{2}$, la moitié de l'effectif total).

Remarque importante : on construit de même les polygones des fréquences cumulées croissantes et décroissantes.

5 LES PARAMETRES DE POSITION (compléments) :

5.1 LE MODE (du mot français la ..mode)

1. Caractère qualitatif ou quantitatif discret
   a. Un mode (le mode quand il est unique) est une valeur de la variable statistique qui a la plus grande fréquence, qui se rencontre le plus fréquemment ; un caractère peut être bi-modal, etc...; dans le cas de deux valeurs successives de fréquence maximale on parle d'intervalle modal.

2. Caractère quantitatif continu
   a. On procède en deux étapes : recherche de la classe modale, puis du mode

   b. On appelle classe modale, une classe dont la fréquence moyenne par unité d'intervalle de classe est la plus élevée, c'est à dire qui a une **densité** maximale ; une classe modale est représentée par un rectangle de hauteur maximale.

   c. Le mode
      i. Méthode
         (a). On se place dans l'hypothèse d'une classe modale unique $[x_1; x_2[$et l'on s'intéresse uniquement à trois classes : la classe modale et les deux classes qui l'encadrent. On note h_1 la hauteur (effectif corrigé dans le cas d'amplitudes inégales) du rectangle précédent la classe modale, h la hauteur du rectangle modal et h_2 la hauteur du rectangle situé immédiatement à droite du rectangle modal. L'idée est que le mode est attiré du côté du rectangle de plus grande densité.

         Si $h_1 = h_2$, le mode sera le centre de la classe modale.

         Si $h_1 \neq h_2$, on note : $k_1 = h - h_1$ et $k_2 = h - h_2$ et le mode est donné par :

$$\boxed{Mo = \frac{k_2 x_1 + k_1 x_2}{k_2 + k_1}}$$

         Le mode est la moyenne pondérée de x_1 et x_2 respectivement affectés des coefficients k_2 et k_1 (attention à l'inversion de l'ordre).

Remarque : si la classe modale est la première classe alors $h_1 = 0$ et si la classe modale est la dernière classe alors $h_2 = 0$.

ii. Exemple :

(a). Le tableau suivant donne la production laitière (en gallons par semaine) de 4912 vaches (ces données datent de 1922)

Quantité de lait (en gallons par semaine)	Nombre de vaches
7.5 à moins de 12.5	123
12.5 à moins de 15.5	726
15.5 à moins de 18.5	1636
18.5 à moins de 20.5	1530
20.5 à moins de 26.5	821
26.5 à 33.5	76

Représenter cette série par un graphique adapté, donner en justifiant votre réponse, la classe modale de cette série et calculer le mode.

Solution (à prolonger..) :

a_i	b_i	b_i -ai	n_i	d_i	h_i
7.5	12.5	5	123	24.60	49.2
12.5	15.5	3	726	242.00	484.0
15.5	18.5	3	1636	545.33	1090.7
18.5	20.5	2	1530	765.00	1530.0
20.5	26.5	6	821	136.83	273.7
26.5	33.5	7	76	10.86	21.7

5.2 LA MEDIANE

Rappel : dans un triangle une médiane est une droite issue d'un sommet et passant par le milieu du côté opposé ; elle partage le triangle en deux triangles de même aire.

1. **Caractère discret**

   a. Nombre d'observations impair ($n = 2p + 1$) : alors il y a un terme central, c'est la médiane. Si par exemple la série comporte 201 observations, la médiane est la 101eme observation de la série ordonnée.

   b. Nombre d'observations pair : il n'y a pas de terme central, donc à priori pas de médiane ; par convention on prend l'intervalle médian constitué par les deux valeurs centrales et la médiane est la moyenne arithmétique de ces deux valeurs. Si par exemple la série comporte 100 observations, aprés l'avoir ordonné, on prend la 50eme et la 51eme valeur et la médiane est leur moyenne arithmétique.

2. **Caractère quantitatif continu**

   a. On procède par interpolation linéaire, pour déterminer la valeur qui correspond à une fréquence cumulée croissante de 50%.

   b. Graphiquement, on peut déterminer la médiane de trois façons :

      i. En déterminant l'intersection du polygone des effectifs cumulés croissants avec la droite d'équation $y = \frac{n}{2}$ (avec la droite d'équation $y = 0.5$ s'il s'agit des fréquences relatives).

      ii. Même chose avec polygone des effectifs cumulés décroissants.

      iii. En traçant les deux polygones cumulés sur le même graphique et en déterminant leur point d'intersection.

   c. **Détermination par le calcul**

      La détermination de la médiane d'une série continue repose sur l'hypothèse que les observations sont réparties uniformément au sein de chaque classe ; son calcul se fait en deux temps :

      i. Localisation de la classe médiane à l'aide des effectifs cumulés croissants .

ii. Calcul de la médiane par interpolation linéaire. Prenons un exemple : le responsable des ressources humaines d'une entreprise a relevé la distribution statistique suivante correspondant à l'ancienneté du personnel cadre dans l'entreprise :

Classes	n_i	$n_i cc$
$[6.5; 8[$	3	3
$[8; 9.5[$	8	11
$[9.5; 11[$	12	23
$[11; 12.5[$	19	42
$[12.5; 14[$	9	51
$[14; 15.5[$	5	56
$[15.5; 17[$	4	60
Total	60	

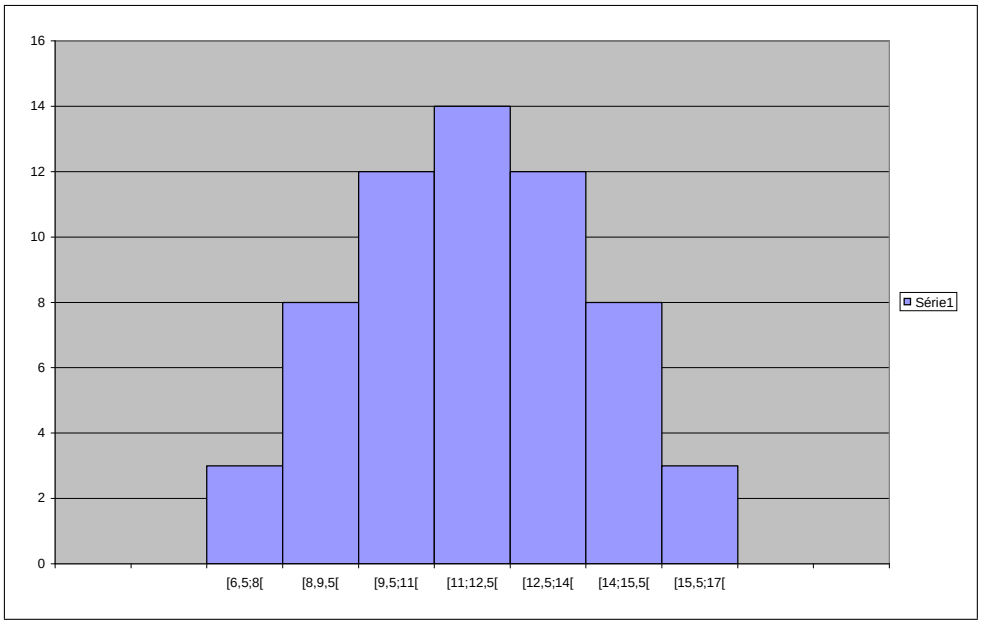
Notons x la médiane (Me).

L'effectif moitié est de 30 et les effectifs cumulés croissants permettent de localiser la médiane dans la classe $[11; 12.5[$; appelons $A\,(11; 23)$ et $B\,(12.5; 42)$ les points extrêmes de la classe médiane et $M\,(x\,; 30)$ le point d'abscisse x ; écrivons l'alignement des points A, B et M, soit par exemple l'égalité des coefficients directeurs des droites (AB) et (AM) : $\dfrac{y_B - y_A}{x_B - x_A} = \dfrac{y_M - y_A}{x_M - x_A}$ soit : $\dfrac{42-23}{12.5-11} = \dfrac{30-23}{x-11} \Leftrightarrow \dfrac{9}{1.5} = \dfrac{7}{x-11}$ soit en écrivant l'égalité des produits en croix : $9\,(x-11) = 7 * 1.5$ soit $x = \dfrac{10.5}{9} + 11 = \dfrac{105}{9} = \dfrac{35}{3} \simeq \mathbf{11.67}$.

Si vous tracez sur le graphique de l'histogramme la droite d'équation $x = 11.7$, vous constaterez que cette droite partage l'histogramme en deux parties de même aire, donc représentant chacune 50% de l'effectif total.

iii. Remarque : dans une distribution symétrique (répartition égale des valeurs autour d'une valeur centrale), la moyenne $\overline{x}$ et la médiane sont égales (le mode aussi). Vérifier le sur cet exemple :

Classes	Effectifs : n_i
$[6.5; 8[$	3
$[8; 9.5[$	8
$[9.5; 11[$	12
$[11; 12.5[$	14
$[12.5; 14[$	12
$[14; 15.5[$	8
$[15.5; 17[$	3
Total	60



Distribution symétrique

5.3 LES QUANTILES (ou fractiles)

Le principe est le même que pour la médiane : il s'agit de partager la distribution **ordonnée** (en ordre croissant) en plusieurs (p parties égales).

1. $p = 2$ donne la médiane.
2. $p = 4$ donne les **QUARTILES.** Il y a trois quartiles, que l'on nomme $Q_1; Q_2; Q_3$; 25% des valeurs de la série sont inférieures à Q_1.

$Q_2 = Me$, car 50% des valeurs sont inférieures à Q_2. Q_3 est le troisième quartile (75% des observations) et $Q_3 - Q_1$ est appelé l'écart interquartile ; l'intervalle interquartile $[Q_1; Q_3[$ comporte 50% des observations.

Le calcul et la détermination graphique des quartiles sont identiques à ceux de la médiane.

	Q_1	Q_2	Q_3
25%	25%	25%	25%

3. $p = 10$ donne **les Déciles.** $D_1; D_2; D_3, ...D_9$ $(10\%; 20\%; .; etc.)$.
4. $p = 100$ donne **les Centiles** $C_1; C_2; C_3, ...C_{99}$ $(1\%; 2\%; .; etc.)$.

5.4 LA MOYENNE ARITHMETIQUE, *notée* $\overline{x}$

1. Moyenne simple (pas de coefficient de pondération) : $\boxed{\overline{x} = \dfrac{\sum x_i}{n}}$
2. Moyenne arithmétique pondérée : $\boxed{\overline{x} = \dfrac{\sum n_i x_i}{n}}$ avec $n = \sum n_i$ ou $\overline{x} = \sum f_i x_i$

   Dans le cas des distributions continues, on prend pour x_i, le centre de classe (du fait de l'hypothèse de répartition uniforme à l'intérieur d'une classe). La moyenne arithmétique pondérée, traditionnellemt notée $\overline{x}$, est donnée par la formule : $\boxed{\overline{x} = \dfrac{\sum n_i x_i}{n} = \sum f_i x_i}$
3. Exemples :

   a.

Nombre d'enfants : x_i	Ménages : n_i	$n_i x_i$
0	50	0
1	60	60
2	40	80
3	20	60
4	5	20
5	5	25
Total	180	245

   i. $\sum n_i x_i = 245$ et $\overline{x} = \dfrac{245}{180} = \dfrac{49}{36} \simeq 1.36$ enfants

   b. Dans le cas d'un caractère continu, pour calculer la moyenne, on remplace chaque classe par le centre de classe : x_i.

   i.

Age	Demandes d'emploi : n_i	x_i	$n_i x_i$
$[20; 25[$	475	22.5	10687.5
$[25; 50[$	355	37.5	13312.5
$[50; 60[$	139	55	7645.0
$[60; 65[$	31	62.5	1937.5
Total	1000		33582.5

   $$\overline{x} = \frac{33582.5}{1000} = 33.58.$$

   c. **Remarques**

   ↬ $\boxed{n\overline{x} = \sum n_i x_i}$ cette formule traduit bien le rôle de la moyenne arithmétique, qui est une moyenne relative à l'addition : on ne change pas la somme des termes d'une série en remplaçant chaque terme par la moyenne arithmétique. Si une série de 100 termes à pour moyenne arithmétique 11.35, on en déduit que la somme de ses termes est de 1135.

   ↬ $\boxed{\overline{x} = \sum f_i x_i}$ cette formule est indispensable dans les exercices où l'on ne connait que les fréquences.

   Exemple : dans une usine de fabrication de chocolat on a constaté en contrôlant le poids de tablettes de 125 g que 40% d'entre elles avaient un poids compris entre 124 et 125g et 60% entre 125 et 126g ; quel est le poids moyen ?

$\overline{x} = 0.40 * 124.5 + 0.60 * 125.5 = 125.1g$

d. **Propriétés**

i. On notera que la moyenne arithmétique est rarement une valeur observée de la série ; elle en constitue le ”centre de gravité”.

ii. La moyenne arithmétique est sensible aux valeurs extrêmes de la série (elle n'est pas robuste) : prenons par exemple la série $\{1; 1; 2; 2; 2; 4; 4; 4; 70\}$, sa moyenne est de 10 et ne constitue pas une valeur centrale représentative de cette série ; par contre la série $\{1; 1; 2; 2; 2; 4; 4; 4\}$ a une moyenne de 2.5 ce qui semble une valeur centrale raisonnable.

iii. On **centre** une série en introduisant la nouvelle série : $y_i = x_i - \overline{x}$ dont la moyenne $\overline{y}$ est nulle. y est une variable statistique centrée.

Centrer la série $\{1; 1; 2; 2; 2; 4; 4; 4\}$.

iv. **Moyennes partielles**

Si nous séparons la série précédente en deux séries $S_1 = \{1; 1; 2; \}$ et $S_2 = \{2; 2; 4; 4; 4\}$, calculer $\overline{x}_1$ et $\overline{x}_2$; comment retrouver $\overline{x}$?

De manière générale si l'on agrège deux séries S_1 et S_2 , on aura : $\boxed{\overline{x} = \dfrac{n_1\overline{x}_1 + n_2\overline{x}_2}{n}}$ avec $n = n_1 + n_2$.

On peut notamment utiliser cette propriété pour mettre à jour une moyenne arithmétique lorsque l'on ajoute une valeur supplémentaire. Appliquer cette formule pour $S_1 = \{1; 1; 2; 2; 2; 4; 4; 4\}$ et $S_2 = \{70\}$.

v. **Changement de variable affine**

changement de variable de type affine : $y = ax + b$;

$$\text{Si } y = ax + b, \text{ alors } \boxed{\overline{y} = a\overline{x} + b}$$

6 PARAMETRES DE DISPERSION

1. **L'ÉTENDUE (l'empan)**

C'est la différence entre la plus grande et la plus petite ; c'est la plus simple des caractéristiques de dispersion, d'un usage courant, elle est souvent utilisée sous la dénomination : éventail ou fourchette *(*éventail des salaires*).*

Ce paramètre possède évidemment le gros désavantage d'être particulièrement sensible aux valeurs extrêmes.

2. **ÉCART MOYEN ABSOLU**

C'est la moyenne arithmétique des écarts (en valeur absolue) entre les valeurs observées et leur moyenne arithmétique.

$\boxed{e_m = \dfrac{1}{n}\sum n_i \left|x_i - \overline{x}\right|}$. Il donne la distance moyenne d'une observation à la valeur centrale.

3. **ÉCART MÉDIAN ABSOLU**

On reprend la même formule, mais on prend comme valeur centrale la médiane (Me) à la place de la moyenne.

$\boxed{e'_m = \dfrac{1}{n}\sum n_i \left|x_i - Me\right|}$

4. **LES ÉCARTS INTERQUANTILES**

On appelle écart interquantile la longueur de l'intervalle constitué par le premier et le dernier quantile. Les plus utilisés sont l'écart interquartile et l'écart interdécile.

	Interquartile	Interdécile
Ecart	$Q_3 - Q_1$	$d_9 - d_1$
Pourcentage d'observations	50%	80%

5. **VARIANCE ET ÉCART-TYPE**

a. **Variance** : C'est la moyenne arithmétique des carrés des écarts entre les valeurs observées et leur moyenne arithmétique.

$\boxed{V(x) = \frac{1}{n}\sum n_i (x_i - \overline{x})^2 = \overline{(x-\overline{x})^2}}$ On utilise plutôt pour les calculs la formule équivalente :

$$\boxed{V(x) = \frac{1}{n}\sum n_i {x_i}^2 - (\overline{x})^2 = \overline{x^2} - (\overline{x})^2} : \text{"}MC - CM\text{"}$$

Moyenne des carrés − Carré de la moyenne. (la variance est toujours positive ou nulle).

b. **Ecart-type** : C'est la mesure de dispersion la plus utilisée ; l'écart-type, noté $\sigma(x)$, est la racine carré de la variance . L'écart-type mesure les fluctuations de la série autour de sa moyenne, les milieux financiers parlent souvent de volatilité d'une action.

$\boxed{\sigma(x) = \sqrt{V(x)}}$

Exemple

On considère un appartement comprenant 5 pièces carrées : 3 pièces de 4m de côté et 2 pièces de 5m de côté.

x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
4	3	12	16	48
5	2	10	25	50
	$N = 5$	$\sum n_i x_i = 22$		$\sum n_i x_i^2 = 98$

on a : $\overline{x} = \frac{22}{5} = 4.4$ m(côté moyen) et $MC = \frac{98}{5} = 19.6$ m^2 (surface moyenne) et $CM = \overline{x}^2 = 4.4^2 = 19.36$ (surface du carré moyen) ; on a $V(x) = 19.6 - 19.36 = 0.24$ et $\sigma(x) = \sqrt{0.24} \simeq 0.49$ m ; on s'intéresse en général aux intervalles du type : $]\overline{x} - k\sigma(x) ; \overline{x} + k\sigma(x)[$ qui mesurent la dispersion du caractère statistique autour de sa moyenne ; pour $k = 1$, l'intervalle $]\overline{x} - \sigma(x) ; \overline{x} + \sigma(x)[$ s'appelle l'intervalle standard.

c. Propriétés de la variance

$V(ax) = a^2 V(x)$ et $V(x+b) = V(x)$ soit

$$\boxed{V(ax+b) = a^2 V(x) \text{ et } \sigma(ax+b) = |a|\,\sigma(x)}$$

d. **Le coefficient de variation**

i. La comparaison directe de deux écart-types peut donner une impression fausse pour comparer la dispersion de deux séries. De plus, l'écart-type dépend de l'unité choisie. C'est pourquoi le coefficient de variation, qui mesure la dispersion relative à la moyenne, est utilisé pour comparer la dispersion de plusieurs séries.

Le coefficient de variation est le rapport noté $CV(x)$ et défini par :

$$\boxed{CV(x) = \frac{\sigma(x)}{\overline{x}}}$$

Ce coefficient s'exprime en pourcentage de la moyenne.

ii. Exemple :

Considérons qu'à la suite d'une étude statistique portant sur le poids x des voyageurs et sur celui y des bagages, une compagnie aérienne ait obtenu les résultats suivants (en kg) :

	x	y
Moyenne	70	15
Ecart-type	8	6

Les coefficients de variation donnent respectivement pour la série des voyageurs $CV(x) = 11,43\%$ et pour la série des bagages $CV(y) = 40\%$.

Alors que l'écart-type de la série des voyageurs est plus grand que celui des bagages ($\sigma(x) > \sigma(y)$) la série des poids des bagages est plus dispersée que celle des poids des voyageurs, car $CV(y) > CV(x)$.

Finalement, le coefficient de variation est un nombre sans dimension, indépendant de l'unité de mesure ; il permet de mesurer la dispersion de séries exprimées en unités ou ordres de grandeur différents et mesure

l'homogénéité des données.

7 La boîte à moustache (Box Plot)

Pour mettre en évidence les quartiles d'une distribution, on peut utiliser "la boîte à moustaches " (Box-Plot), qui nous indique quelques traits marquants de la série. On construit une boîte dont les extrémités sont les quartiles Q_1 et Q_3, la longueur de la boîte étant l'écart inter quartile, noté $EIQ = Q_3 - Q_1$; cette boîte est coupée en deux par la médiane Q_2 ; cette boîte contient donc 50% de l'ensemble des observations (25% à gauche de la médiane et 25% à droite). Cette boîte est ensuite prolongée par des moustaches délimitées à gauche par la plus petite valeur de la série et à droite par la plus grande. $\min x_i \longleftarrow$ $\longrightarrow \max x_i$.

En fait on a un problème si la série comporte des valeurs aberrantes, très grandes ou très petites ; pour contourner cet inconvénient et éviter de trop grandes moustaches, on peut utiliser une des deux conventions suivantes :

1. Convention 1 : on limite la taille maximale d'une moustache à $1.5EIQ$, on introduit donc les valeurs pivots qui sont situées de part et d'autre de la boîte et en sont distantes d'une fois et demi sa longueur : ces valeurs sont donc : $a_1 = Q_1 - 1.5\,(Q_3 - Q_1) = Q_1 - 1.5 * EIQ$ et $a_2 = Q_3 + 1.5\,(Q_3 - Q_1) = Q_3 + 1.5 * EIQ$ et on prend pour délimiter les moustaches les valeurs extrêmes de la série situées dans cet intervalle : à gauche $x_g = \min\{x_i/x_i \geq Q_1 - 1.5 * EIQ\}$et à droite $x_d = \max\{x_i/x_i \leq Q_3 + 1.5 * EIQ\}$

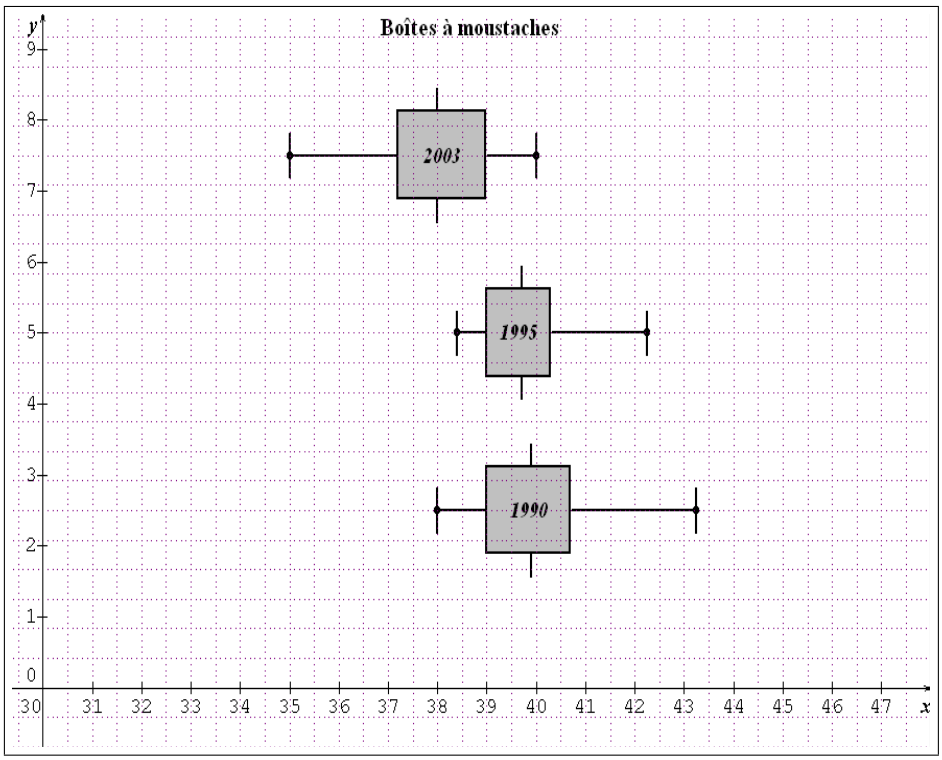
   On constate que la plupart des séries qui ne contiennent pas de valeurs aberrantes se situent dans l'intervalle $[a_1; a_2]$.

2. Convention 2 : on peut aussi utiliser les déciles pour représenter les moustaches, la moustache de gauche commence au premier décile d_1 et celle de droite finit à d_9,on prend donc en compte 80% de la série, on écarte les 10% valeures plus faibles et les 10% plus grandes.

Ce type de diagramme permet de comparer facilement plusieurs distributions.

Exemple : on considère le tableau suivant donnant la durée hebdomadaire du travail des salariés à temps complet dans certains pays de l'union européenne.

Pays	**1990**	**1995**	2003
Allemagne	39.9	39.7	37,7
Autriche	40.1	39.3	38,5
Belgique	38	38.4	38
Danemark	39	39	37
Espagne	40.7	40.7	38,6
Finlande	38.4	38.6	37,5
France	39.6	39.9	35
Grèce	40.2	40.3	40
Irlande	40.4	40.2	39
Italie	38.6	38.4	38
luxembourg	39.9	39.5	39
Pays -Bas	39	39.5	37
portugal	41.9	41.2	39
Royaume -Uni	43.7	43.9	37,2
Suède	40.7	40	38,8



	1990	1995	2003
1	38	38,4	35
2	38,4	38,4	37
3	38,6	38,6	37
4	39	39	37,2
5	39	39,3	37,5
6	39,6	39,5	37,7
7	39,9	39,5	38
8	39,9	39,7	38
9	40,1	39,9	38,5
10	40,2	40	38,6
11	40,4	40,2	38,8
12	40,7	40,3	39
13	40,7	40,7	39
14	41,9	41,2	39
15	43,7	43,9	40

	1990	1995	2003
Q_1	39	39	37,2
Q_2	39,9	39,7	38
Q_3	40,7	40,3	39
EIQ	1,7	1,3	1,8
1,5EIQ	2,55	1,95	2,7
Q_1-1,5EIQ	36,45	37,05	34,5
Q_3+1,5EIQ	43,25	42,25	41,7

On a représenté les boîtes à moustaches des distributions 1990, de 1995 et de 2003, après avoir ordonné les séries et effectué les calculs nécessaires.