Université de Picardie Jules Verne **2015-2016**
UFR des Sciences

Licence Sciences, Technologies, Santé toutes mentions - Semestre 2
Probabilités et Statistique

Statistique descriptive à une variable

L'objectif est de présenter les notions essentielles de la statistique descriptive, c'est-à-dire à de montrer comment décrire de façon claire et concise l'information apportée par des observations nombreuses et variées sur un phénomène donné. Il s'agit de trier ces données, les décrire, les résumer sous forme de tableaux, de graphiques, et sous forme d'un petit nombre de paramètres-clés (moyenne, médiane par exemple).

1. Différents types de variables

On considère les individus d'une population donnée et une variable X décrivant ces individus.

1.1. Variables quantitatives

Ce sont des variables qui prennent des valeurs numériques. Elles sont de **deux types**.

Les **variables discrètes** ne prennent leurs valeurs que dans un ensemble de valeurs disjointes (par exemple des nombres entiers) ; autrement dit, aucune valeur n'est prise dans l'intervalle séparant deux valeurs possibles consécutives. Par exemple, le nombre d'individus d'une espèce par hectare, le nombre de vertèbres d'une larve de poisson, ...

Les **variables continues** peuvent à priori prendre toutes les valeurs d'un intervalle de variation (par exemple $[-2;5]$ ou $]-\infty;+\infty[$). C'est le cas de la plupart des mesures de longueurs, surfaces, volumes, masses, concentrations, températures, ...

On désignera par x_i les valeurs observées sur un échantillon ; elles seront naturellement rangées dans l'ordre croissant.

1.2. Variables qualitatives.

Elles définissent une propriété non quantifiable. Par exemple, une couleur, un génotype, un phénotype, l'appartenance à une espèce ou à une variété, ... Au lieu des valeurs, on parlera des modalités d'une telle variable ; on les désignera encore par x_i. Une variable **nominale** est une variable qualitative dont les modalités ne sont pas ordonnées ; par exemple la couleur des yeux (bleus, verts, noirs, ...)

Elles peuvent elles aussi être discrètes ou continues. Par exemple, les génotypes ou les espèces sont discrètes : elles n'admettent pas de modalité intermédiaire. Par contre, pour les couleurs non spectrales, tout intermédiaire est concevable à l'intérieur d'une plage de variation (comme l'échelle de Forel adoptée pour les couleurs des eaux naturelles allant du brun au bleu en passant par le vert).

1.3. Variables semi-quantitative ou ordinale.

A défaut d'être mesurables, elles sont classables.

Ou bien les modalités d'une variable qualitative se succèdent dans un ordre naturel ; par exemple les stades de développement d'un organisme, taille d'un vêtement (XS, S, M, L, XL, XXL)

Ou bien les valeurs observées d'une variable quantitative sont classées par ordre croissant ou décroissant, et on néglige leur valeur précise pour ne retenir que leur rang dans ce classement. Ce rang est alors une variable discrète mais ne correspond pas à une mesure.

1.4. Le regroupement en classes.

Très souvent, la prise en compte de toute la précision d'une mesure (et donc de toute les valeurs observées) ne permet pas de donner une interprétation simple des résultats et conduit à des calculs inutiles. On peut souvent se contenter de regarder des classes de valeurs C_i, c'est-à-dire des intervalles d'amplitudes fixées ; des mesures tombant dans une même classe devenant alors indiscernables. C'est ce que l'on fait en particulier pour les variables quantitatives continues, les clases étant naturellement rangées dans l'ordre croissant. Par exemple, si on mesure la taille des individus d'un échantillon de la population française, il est inutile d'avoir le nombre exact d'individus mesurant chaque taille (160cm, 161cm, 162cm, ...), mais suffisant de savoir combien mesurent entre 160 et 165 cm, entre 165 et 170 cm, ... Mais on perd de l'information ...

2. Représentation d'une variable

Dans une population, on considère un échantillon de n individus sur lequel on observe une variable X.
Si X est quantitative discrète, on parlera des valeurs x_i de la variable X.
Si X est qualitative discrète, on parlera des modalités x_i de la variable X.
Si X est quantitative continue, on parlera des classes C_i de la variable X.

2.1. Distributions des fréquences ; diagramme circulaire ou en barres, diagramme en bâtons ou histogramme

Le nombre d'individus étant généralement grand, voire très grand, une série brute est difficilement lisible et interprétable. Il est indispensable de la résumer. Pour cela, on commence par un **tri à plat**, décompte des modalités ou valeurs obtenues, qui nous servira de base à la construction de tableaux et de graphiques.

On désigne par n_i l'**effectif** (ou fréquence absolue) de chaque valeur ou modalité x_i ou de chaque classe C_i, c'est-à-dire le nombre de fois où l'on a observé la valeur ou la modalité ou la classe dans l'échantillon.

On a évidemment $\sum n_i = n$.

On désigne par $f_i = \frac{n_i}{n}$ la **fréquence** (ou fréquence relative ou proportion) de chaque valeur ou modalité x_i ou de chaque classe C_i, c'est-à-dire la proportion de fois où l'on a observé la valeur ou la modalité ou la classe dans l'échantillon. (Pourcentage = $100 \times f_i$).

On peut remarquer que $\sum f_i = \sum \frac{n_i}{n} = \frac{1}{n}\sum n_i = \frac{1}{n} n = 1$.

La **distribution des effectifs et/ou fréquences** est en général donnée comme suit :

Valeur ou modalité x_i	Effectif n_i	Fréquence f_i
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$		
x_p	n_p	f_p
	n	1

Classe C_i	Effectif n_i	Fréquence f_i
C_1	n_1	f_1
C_2	n_2	f_2
$\vdots$		
C_p	n_p	f_p
	n	1

Lorsque l'on veut comparer les résultats de plusieurs échantillons (éventuellement de tailles différentes), il est utile d'utiliser les fréquences. C'est ce que nous ferons en général.

La représentation graphique d'une telle distribution est de différents types :

- **diagramme circulaires** (ou camembert) pour une variable qualitative : chaque modalité est représentée par un secteur circulaire dont l'angle (et donc la surface) est proportionnel à son effectif ou sa fréquence. Le rayon du cercle est arbitraire.
- **diagramme en barres** pour une variable qualitative : chaque modalité est représentée par un rectangle de base constante et d'une hauteur proportionnelle à son effectif ou à sa fréquence.
- **diagramme en bâtons** pour une variable quantitative discrète : à chaque valeur x_i portée en abscisse on fait correspondre un bâton ayant pour hauteur la fréquence f_i (portée en ordonnée) ; la somme des hauteurs étant égale à 1. Le **polygone des fréquences** est la ligne brisée obtenue en reliant les sommets des bâtons ; il sert à guider l'oeil afin d'avoir une vision globale du diagramme.
- **histogramme** pour une variable quantitative continue avec regroupement en classes C_i : à chaque classe C_i d'amplitude a_i on fait correspondre un rectangle de hauteur $h_i = \frac{f_i}{a_i}$; ainsi, l'aire des rectangles est proportionnelle à la fréquence ; la somme des aires étant égale à 1.

2.2. Courbe des fréquences cumulées.

Cas des valeurs ou modalités x_i.

A chaque x_i, faisons correspondre la fréquence F_i des valeurs observées inférieures ou égales à x_i. On a alors $F_i = f_1 + \cdots + f_i$ et on dit que F_i est la **fréquence cumulée** correspondant à x_i. Le **diagramme des fréquences cumulées**, construit à partir des points de coordonnées (x_i, F_i), est alors en escalier.

On peut aussi définir les **effectifs cumulés** $N_i = n_1 + \cdots + n_i$, nombre de valeurs observées inférieures ou égales à x_i. On a évidemment $F_i = \frac{N_i}{n}$

Cas des classes $C_i =]\widetilde{x}_{i-1}, \widetilde{x}_i]$. (Remarquer que C_i est fermée à droite et ouverte à gauche.)

Dans ce cas, la **fréquence cumulée** $F_i = f_1 + \cdots + f_i$ correspond à la fréquence des valeurs observées inférieures ou égales à $\widetilde{x}_i$.

Le **polygone des fréquences cumulées** est la ligne brisée obtenue en reliant les points de coordonnées $(\widetilde{x}_i, F_i)$. On écrira alors $F(x) = Fr(X \leq x)$ = fréquence des valeurs inférieures ou égale à x.

Par interpolation, cette courbe permet ainsi de déterminer :

- pour une valeur x la fréquence $F(x)$ des valeurs observées inférieures ou égale à x ;
- pour une fréquence F la valeur x telle que l'on observe des valeurs inférieures ou égales à x avec une fréquence F (voir médiane et quantiles).

3. Paramètres de tendance centrale et paramètres de position

Les paramètres statistiques ont pour but de résumer, à partir de quelques nombres clés, l'essentiel de l'information relative à l'observation d'une variable quantitative. Parmi eux, les paramètres de tendance centrale représentent une valeur numérique autour de laquelle les observations sont réparties (le mode, la médiane, la moyenne), et les paramètres de position (les fractiles).

3.1. Le mode

Le **mode** est la valeur ou la modalité ou la classe ayant la plus grande fréquence d'observation. Dans le dernier cas, on parle aussi de classe modale.

3.2. La médiane

Si la série brute des valeurs observées $x_1, x_2, \ldots, x_n$ est triée par ordre croissant : $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, la médiane M est la valeur du milieu, telle qu'il y ait autant d'observations "au-dessous" que "au-dessus", c'est-à-dire que :

- si n est impair, soit $n = 2p + 1$, alors $M = x_{(p+1)}$;
- si n est pair, soit $n = 2p$, alors toute valeur de l'intervalle médian $[x_{(p)}; x_{(p+1)}]$ répond à la question ;

afin de définir M de façon unique, on choisit souvent $M = \dfrac{x_{(p)} + x_{(p+1)}}{2}$, soit le centre de l'intervalle médian.

Par exemple :

- dans la série $2, 3, 7, 9, 11$, il y a $n = 5 = 2p + 1$ valeurs donc $p = 2$ et on prendra la $(p+1)$-ème $= 3^{ème}$ valeur, soit $M = 7$; 3 valeurs lui sont inférieures ou égales et 3 supérieures ou égales ;
- dans la série $2, 3, 6, 7$, il y a $n = 4 = 2p$ valeurs donc $p = 2$ et on pourra prendre $M = \dfrac{x_{(2)} + x_{(3)}}{2} = \dfrac{3+6}{2} = 4,5$ (mais on pourrait prendre toute valeur de l'intervalle $[3, 6]$).

De manière générale, on définira la **médiane** M comme toute valeur telle qu'au moins 50% des observations lui sont inférieures ou égales et au moins 50% des observations lui sont supérieures ou égales. Dans le cas où il y aurait un intervalle borné non-vide de valeurs répondant à ces deux contraintes, une pratique courante désigne le centre de cet intervalle comme unique médiane.

Pour une variable quantitative continue, on définira la **médiane** M comme étant la valeur (abscisse) correspondant à la fréquence cumulée $F = 0,5$. On l'obtiendra en général par lecture graphique (valeur approchée) sur la courbe des fréquences cumulées, ou par une formule d'interpolation linéaire (valeur exacte).

3.3. Les quantiles

On appelle **quantiles** (**ou fractiles**) **d'ordre k** les valeurs $Q_1, Q_2, \ldots, Q_{k-1}$ qui divisent une série de données ordonnée en k parties d'effectifs égaux. Parmi les fractiles, on trouve :

- la médiane M, valeur dépassant 50% des valeurs observées ;
- les quartiles Q_1, Q_2 et Q_3, valeurs dépassant respectivement 25%, 50% et 75% des valeurs observées (remarquer que $Q_2 = M$) ;
- les déciles D_1, D_2, ..., D_9, valeurs dépassant respectivement 10%, 20%, ..., 90% des valeurs observées ;
- les centiles ...

Comme la médiane, les quantiles peuvent être obtenus graphiquement à l'aide de la courbe des fréquences cumulées, ou par une formule d'interpolation linéaire.

3.4. La moyenne arithmétique

Cas des valeurs x_i avec les effectifs n_i

On désigne par $\bar{x} = \frac{1}{n}\sum n_i x_i$ la **moyenne** arithmétique des valeurs observées. En effet, $\sum n_i x_i$ représente la somme des n valeurs observées.

On peut remarquer que $\bar{x} = \sum \frac{n_i x_i}{n} = \sum f_i x_i$.

Dans le cas d'une série de données "brute" $x_1, x_2, \ldots, x_n$ (certains x_i pouvant être égaux), tous les effectifs n_i sont égaux à 1 et on retrouve $\bar{x} = \frac{1}{n}\sum x_i$.

Cas des classes C_i.

On décide d'appliquer la même formule, dans laquelle x_i représente les centres des classes C_i. On fait comme si on avait observé n_i fois le centre de la classe, ce qui revient à supposer que les valeurs observées dans une classe se répartissent uniformément (ce qui n'est pas forcément le cas).

4. Paramètres de dispersion

Deux distributions peuvent, tout en ayant des caractéristiques de tendance centrale voisines, être très différentes. Ainsi la moyenne annuelle des températures dans une zone océanique peut être égale à la moyenne annuelle des températures dans une zone continentale, pourtant les distributions sont très différentes. Dans le premier cas les variations de température autour de la moyenne sont assez faibles, dans le second cas elles sont beaucoup plus importantes.

Il est donc nécessaire de mesurer la dispersion des valeurs autour des tendances centrales.

4.1. L'étendue

L'**étendue** (ou amplitude), notée R (Range), d'une série statistique est la différence entre la valeur maximum et la valeur minimum de la série. Facile à déterminer, l'étendue ne dépend que des 2 observations extrêmes qui sont parfois le fait de situations exceptionnelles. Il est donc difficile de considérer l'étendue comme une mesure stable de la dispersion.

4.2. L'écart interquartile

Afin de diminuer l'influence des valeurs extrêmes on peut tenir compte de valeurs plus stables de la distribution. Par exemple, l'intervalle interquartile $[Q_1, Q_3]$ ou l'**écart interquartile** $EIQ = Q_3 - Q_1$ mesure la dispersion des valeurs observées autour de la médiane.

4.3. Quel écart à la moyenne ?

On veut savoir si les valeurs observées s'écartent beaucoup de la moyenne.

On peut naturellement considérer la **moyenne des écarts à la moyenne**, c'est-à-dire

$$\frac{1}{n}\sum n_i(x_i - \bar{x}) = \frac{1}{n}\left(\sum n_i x_i - \sum n_i \bar{x}\right) = \frac{1}{n}\left(n\bar{x} - \bar{x}\sum n_i\right) = \frac{1}{n}(n\bar{x} - \bar{x}n) = 0 \text{ !!!}$$

En fait, les écarts positifs et négatifs sont compensent exactement.

On pourrait alors considérer la **moyenne des valeurs absolues des écarts à la moyenne**, c'est-à-dire

$$\frac{1}{n}\sum n_i|x_i - \bar{x}|.$$

Ce paramètre serait tout à fait acceptable mais pour des raisons mathématiques (calculs de probabilités), on lui préfère la **moyenne des carrés des écarts à la moyenne**, c'est-à-dire la variance.

4.4. La variance et l'écart-type.

On désigne par $var(x) = s_x^2 = \frac{1}{n}\sum n_i(x_i - \bar{x})^2$ la **variance** des valeurs observées.

Dans cette formule, comme pour la moyenne, x_i représente les valeurs observées ou les centres des classes C_i.

On peut démontrer que $s_x^2 = \frac{1}{n}\sum n_i x_i^2 - \bar{x}^2$ (formule plus rapide pour les calculs).

Dans le cas d'une série de données "brute" $x_1, x_2, \ldots, x_n$ (certains x_i pouvant être égaux), tous les effectifs n_i sont égaux à 1 et on retrouve $s_x^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$.

On désigne par $s_x = \sqrt{s_x^2} = \sqrt{var(x)}$ l'**écart-type** des valeurs observées.

On désigne par $\left[\bar{x} - s_x\ ;\ \bar{x} + s_x\right]$ l'**intervalle moyen**. On dit qu'en moyenne, les valeurs observées se trouvent dans l'intervalle moyen.

5. D'autres paramètres

5.1. Le coefficient de variation

Le **coefficient de variation** est le rapport de l'écart-type à la moyenne : $V = \frac{s_x}{\bar{x}}$. On peut l'exprimer en pourcentage en le multipliant par 100.

Il donne une mesure relative de l'écart type qui permet de prendre en compte l'ordre de grandeur de la moyenne.

5.2. Les coefficients de forme

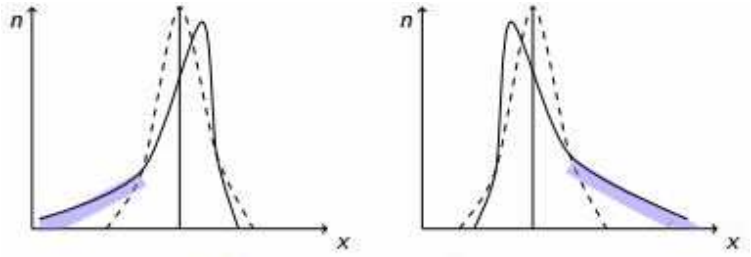
On définit d'autres coefficients pour caractériser la forme du diagramme des effectifs.

Le **coefficient d'asymétrie** ou Skewness est la moyenne des cubes des valeurs centrées des observations : $\mu_3 = \frac{1}{n}\sum n_i(x_i - \bar{x})^3$.

Le **coefficient d'asymétrie** ou Skewness de Fisher (relatif) est la moyenne des cubes des valeurs centrées réduites des observations : $S = \frac{1}{n}\sum n_i\left(\frac{x_i - \bar{x}}{s_x}\right)^3 = \frac{\mu_3}{s_x^3}$.

Interprétation :

- lorsque la distribution est symétrique, le coefficient de Skewness est nul ;
- lorsque la distribution possède une forte queue vers la droite, le coefficient de Skewness est positif (les + l'emportent) ;
- lorsque la distribution possède une forte queue vers la gauche, le coefficient de Skewness est négatif (les - l'emportent).



Skewness à gauche, $S < 0$ [& mode à droite de la moyenne]

Skewness à droite, $S > 0$ [& mode à gauche de la moyenne]

Le **coefficient d'aplatissement** ou Kurtosis de Pearson est la moyenne des puissances quatrièmes des observations centrées : $\mu_4 = \frac{1}{n}\sum n_i(x_i - \bar{x})^4$.

Le **coefficient d'aplatissement** ou Kurtosis de Pearson (relatif) est la moyenne des puissances quatrièmes des observations centrées réduites : $K = \frac{1}{n}\sum n_i\left(\frac{x_i - \bar{x}}{s_x}\right)^4 = \frac{\mu_4}{s_x^4}$.

Interprétation : il permet d'etudier la forme plus ou moins pointue ou aplatie du diagramme des effectifs.

Fisher propose d'étudier $K' = K - 3$, ce qui permet de faire référence à une distribution particulière, celle de la loi normale (ou gaussienne, étudiée dans le dernier chapitre de ce cours) pour laquelle K vaut 3.

Les logiciels statistiques vous donnent la valeur de K'. Si $K' > 0$, alors la distribution est plus aplatie que dans une distribution normale ; si $K' < 0$, alors elle l'est moins.

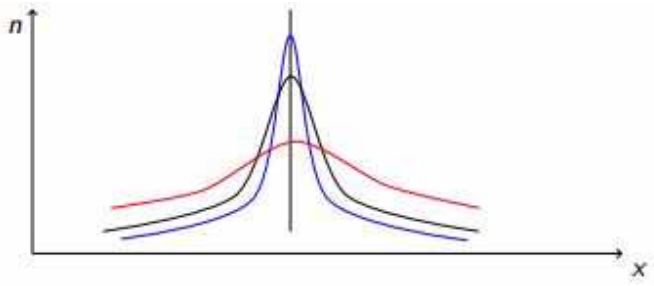


FIGURE: Un kurtosis positif ($K' > 0$) indique que les *queues* comptent *plus d'observations* que dans une distribution gaussienne. Un kurtosis négatif ($K' < 0$) indique que les *queues* comptent *moins d'observations* que dans une distribution gaussienne. Un kurtosis nul est celui d'une loi gaussienne

6. Exemples de traitement statistique d'observations sur un échantillon

6.1. Cas discret

Pour étudier le nombre d'enfants dans les familles amiénoises, on a interrogé 1000 familles et on a obtenu les résultats suivants :

Nombre d'enfants	0	1	2	3	4	5	6	7	8
Nombre de familles	162	240	297	208	62	16	8	5	2

On peut considérer la situation suivante.
Population : les familles.
Variable X : le nombre d'enfants, quantitative discrète.
Echantillon de $n = 1000$ familles.

Valeur x_i	Effectif n_i	Eff. Cum. N_i	Fréquence f_i	Fréq. Cum. F_i	$n_i x_i$	$n_i x_i^2$
0	162	162	$0,162$	$0,162$	0	0
1	240	402	$0,240$	$0,402$	240	240
2	297	699	$0,297$	$0,699$	594	1188
3	208	907	$0,208$	$0,907$	624	1872
4	62	969	$0,062$	$0,969$	248	992
5	16	985	$0,016$	$0,985$	80	400
6	8	993	$0,008$	$0,993$	48	288
7	5	998	$0,005$	$0,998$	35	245
8	2	1000	$0,002$	1	16	128
Total	1000		1		1885	5353

Mode : 2 ; c'est la valeur la plus fréquemment observée.

Médiane : $M = \frac{x_{(500)} + x_{(501)}}{2} = \frac{2+2}{2} = 2$: au moins 50% des familles ont un nombre d'enfants inférieur ou égal à 2, et au moins 50% des familles ont un nombre d'enfants supérieur ou égal à 2.

Quartiles : $Q_1 = \frac{x_{(250)} + x_{(251)}}{2} = \frac{1+1}{2} = 1$, $Q_2 = M = 2$ et $Q_3 = \frac{x_{(750)} + x_{(751)}}{2} = \frac{3+3}{2} = 3$.

Moyenne : $\bar{x} = \frac{1}{n}\sum n_i x_i = \frac{1}{1000} \times 1885 = 1,885$.

Variance : $s_x^2 = \frac{1}{n}\sum n_i x_i^2 - \bar{x}^2 = \frac{1}{1000} \times 5353 - (1.885)^2 \simeq 1,7998$.

Ecart-type : $s_x = \sqrt{s_x^2} \simeq 1,3416$.

Intervalle moyen : $\left[\bar{x} - s_x\,;\, \bar{x} + s_x\right] = \left[0,54\,;\, 3,23\right]$; en moyenne, les familles ont un nombre d'enfants compris entre 1 et 3 enfants..

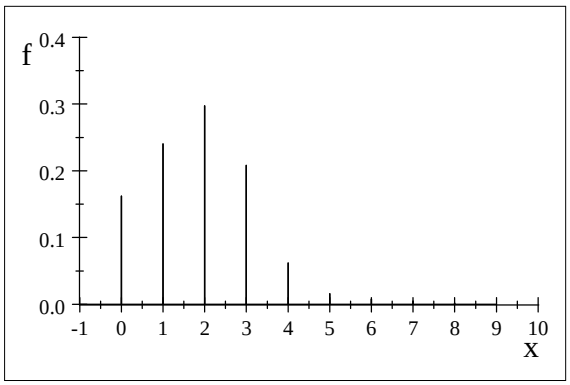


Diagramme en bâtons des fréquences

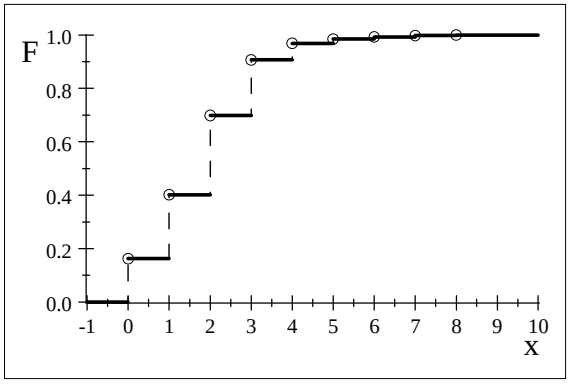


Diagramme des fréquences cumulées

6.2. Cas continu

Un échantillon de 50 poissons de la même espèce à fourni les poids suivants (en g) :

61	82	92	97	101	104	109	118	131	155
69	82	93	97	101	104	110	120	133	165
70	85	93	99	101	105	110	121	138	166
74	85	93	99	102	106	110	125	140	180
79	87	94	99	102	107	114	128	147	180

On peut considérer la situation suivante.

Population : les poissons d'une espèce donnée.

Variable X : le poids, quantitative continue (même si observations arrondies à l'entier le plus proche).

Echantillon de $n = 50$ poissons.

Presque toutes les valeurs n'étant observées qu'une fois, une étude analogue à celle de l'exemple 1 ne permettrait pas de résumer l'information de façon significative. On préfèrera donc regrouper les valeurs en classes de poids. On aurait pu découper l'intervalle de variation des valeurs $[60;190]$ en classes de même amplitude (10g par exemple) mais certaines classes (les premières et dernières classes) auraient eu des effectifs (et donc des fréquences) très faibles. C'est pourquoi on a choisi des classes d'amplitude plus grande en début et fin de distribution.

Classe C_i	Centre x_i	Effectif n_i	Fréq. f_i	Ampl. a_i	Haut. h_i	Freq. Cum. F_i	$n_i x_i$	$n_i x_i^2$
$[60;80]$	70	5	$0,10$	20	$0,005$	$0,10$	350	24500
$]80;90]$	85	5	$0,10$	10	$0,010$	$0,20$	425	36125
$]90;100]$	95	10	$0,20$	10	$0,020$	$0,40$	950	90250
$]100;110]$	105	14	$0,28$	10	$0,028$	$0,68$	1470	154350
$]110;120]$	115	3	$0,06$	10	$0,006$	$0,74$	345	39675
$]120;130]$	125	3	$0,06$	10	$0,006$	$0,80$	375	46875
$]130;140]$	135	4	$0,08$	10	$0,008$	$0,88$	540	72900
$]140;180]$	160	6	$0,12$	40	$0,003$	1	960	153600
		50	1				5415	618275

Classe modale : $]100;110]$; c'est la classe la plus fréquemment observée.

Médiane : $M = 100 + (110 - 100)\dfrac{0,50 - 0,40}{0,68 - 0,40} \simeq 103,6$; la moitié des poissons ont un poids inférieur ou égal à 103 g, l'autre moitié ayant un poids supérieur à 103 g.

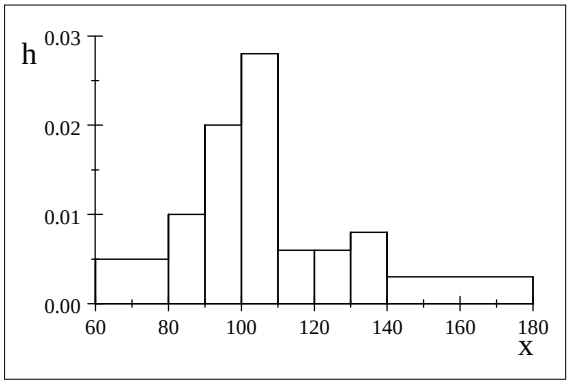
Quartiles : $Q_1 = 90 + (100 - 90) \times \dfrac{0,25 - 0,20}{0,40 - 0,20} = 92,5$ et $Q_3 = \cdots \simeq 121,7$.

Moyenne : $\bar{x} = \dfrac{1}{n}\sum n_i x_i = \dfrac{1}{50} \times 5415 = 108,3$.

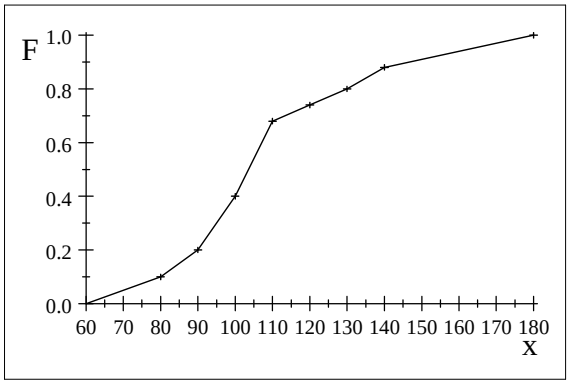
Variance : $s_x^2 = \dfrac{1}{n}\sum n_i x_i^2 - \bar{x}^2 = \dfrac{1}{50} \times 618275 - \left(\dfrac{1}{50} \times 5415\right)^2 = 636,61$.

Ecart-type : $s_x = \sqrt{s_x^2} \simeq 25,2$.

Intervalle moyen : $\left[\bar{x} - s_x \,;\, \bar{x} + s_x\right] = \left[83,1 \,;\, 133,5\right]$.



Histogramme des fréquences



Polygone des fréquences cumulées

7. Exercices

Exercice 1.

On a interrogé des personnes au hasard et on a obtenu les résultats suivants

Personne n°	Nom	Prénom	Age	Salarié	Niveau étude	Département de naissance	Sexe
1	PASCAL	Béatrice	22	Non	Primaire	80	F
2	NOIROT	Claudine	25	Oui	Universitaire	78	F
3	LAFFITE	Jean	30	Oui	Secondaire	93	M
4	LAFFON	Gilles	25	Non	Primaire	80	M
5	DAURIAC	André	30	Oui	Universitaire	32	M
6	FAURE	Céline	22	Non	Universitaire	64	F
7	BENAZET	Eric	24	Non	Secondaire	40	M
8	DUMAS	Elvia	29	Non	Secondaire	76	F
9	MARTINEZ	Alexis	25	Oui	Universitaire	80	M
10	DUPONT	Adrien	23	Non	Universitaire	75	M
11	CATHALA	Agnès	22	Non	Primaire	78	F
12	PEREZ	Eliane	24	Oui	Secondaire	13	F
13	MARTIN	Albert	25	Oui	Universitaire	33	M
14	SIMON	Gabriel	24	Oui	Primaire	76	M
15	ROQUES	Adrien	25	Non	Secondaire	45	M
16	DUMAS	Elvire	28	Oui	Secondaire	75	F
17	MARTIN	Alain	25	Oui	Secondaire	21	M
18	SANCHEZ	Henri	27	Oui	Primaire	11	M
19	PONS	Adeline	22	Non	Universitaire	34	F
20	DUPUY	Paul	27	Oui	Universitaire	73	M

1) Combien y a-t-il d'individus ? de variables ?
2) Pour chaque variable, préciser sa nature (autrement dit son type).
3) Représenter les résultats sur le Niveau d'étude à l'aide d'un diagramme circulaire.
4) Représenter les résultats sur le Niveau d'étude à l'aide d'un diagramme en barres.
5) a) Représenter les résultats sur l'Age à l'aide d'un diagramme adapté.
   b) Déterminer le mode, la moyenne et la médiane de l'Age des individus étudiés.

Exercice 2.

Cinquante éprouvettes d'acier spécial sont soumises à des essais de résistance. Pour chacune, on note le nombre de chocs nécessaires pour obtenir la rupture. Les résultats obtenus sont les suivants :

2	2	3	5	2	1	4	2	3	5
3	2	3	3	4	1	2	4	2	2
4	2	3	2	3	3	2	2	4	2
1	4	2	3	2	2	3	1	3	3
2	3	2	2	3	4	3	2	3	2

1) Préciser la population étudiée, la variable étudiée et sa nature, la taille de l'échantillon.
2) Représenter ces résultats sous forme d'un tableau valeurs/effectifs (tri à plat).
3) Tracer sur le même graphique le diagramme et le polygone des fréquences de cette distribution. En déduire le mode et donner sa signification.
4) Tracer le diagramme des fréquences cumulées (croissantes). Déterminer les quartiles.
5) Déterminer la moyenne et l'écart-type de cette série statistique.

Exercice 3. *(D'après partiel de mars 2008)*

Avant d'acheter le dernier modèle d'appareil photo numérique d'une grande marque, un internaute consulte un site web comparateur de prix. L'observation du prix de cet appareil photo chez différents sites marchands donne les résultats indiqués dans le tableau ci-dessous :

Prix (en €)	]500;550]	]550;600]	]600;650]	]650;700]	]700;800]	]800;850]	]850;950]
Nombre de sites	8	8	16	8	8	12	12

1) Préciser la population étudiée, la variable étudiée et sa nature, la taille de l'échantillon.

2) Représenter graphiquement les résultats présentés dans le tableau.

3) Calculer les fréquences cumulées de la distribution et tracer le polygone des fréquences cumulées.

4) En déduire par lecture graphique, puis par une formule d'interpolation linéaire, la valeur de la médiane et des quartiles de la distribution. Interpréter les résultats obtenus.

5) Calculer la moyenne et l'écart-type de la distribution. Préciser les données à partir desquelles ces valeurs ont été calculées.

Exercice 4.

Le croisement d'une souris noire et d'une souris blanche donne des descendants de couleur noire ou blanche. On a effectué 30 croisements ayant donné chacun 50 descendants. Pour chacun des 30 croisements, le nombre de descendants noirs obtenu est donné dans le tableau suivant :

24	28	25	24	26	21	23	21	25	26	18	25	26	29	25
22	25	26	32	25	23	24	25	25	27	29	19	24	27	26

1) Calculer la moyenne et l'écart-type de cette série. Déterminer la médiane.

2) Ranger ces données en classes d'intervalles de longueur 2, borne supérieure incluse, entre 18 et 32. Faire un tableau.

3) Tracer sur le même graphique l'histogramme et le polygone des fréquences de cette distribution. En déduire la classe modale.

4) Tracer le polygone des fréquences (ou effectifs) cumulées croissantes. En déduire la valeur de la médiane et l'intepréter dans le contexte étudié.

5) Déterminer la moyenne et l'écart-type de cette série statistique. Comparer avec les résultats du 1).

6) Si nous avions rangé ces données en classes d'intervalles de longueur 4, aurions-nous trouvé les mêmes résultats ?

Exercice 5.

On admet le résultat suivant, inégalité de Bienaymé-Tchebichev (version statistique descriptive).

Pour tout réel $t > 0$, la fréquence des observations se trouvant dans l'intervalle $\left[\bar{x} - t \times s_x \,;\, \bar{x} + t \times s_x\right]$ *est au moins égale à* $1 - \frac{1}{t^2}$.

Autrement dit, au moins $\left(1 - \frac{1}{t^2}\right) \times 100$ % des observations se trouvent dans l'intervalle $\left[\bar{x} - t \times s_x \,;\, \bar{x} + t \times s_x\right]$. En pratique :

- soit on se donne t, et alors on obtient le pourcentage d'observations dans l'intervalle correspondant à t : par exemple, pour $t = 2$, on a $1 - \frac{1}{t^2} = \frac{3}{4} = 0.75$, et donc au moins 75 % des observations dans l'intervalle $\left[\bar{x} - 2s_x \,;\, \bar{x} + 2s_x\right]$;

- soit on se donne un pourcentage souhaité, et alors on obtient l'intervalle recherché : par exemple, pour avoir au moins 95 % des observations dans l'intervalle, on cherche t tel que $1 - \frac{1}{t^2} = 0.95$, ce qui donne $t = \sqrt{\frac{1}{0.05}} = \sqrt{20}$, et donc l'intervalle $\left[\bar{x} - \sqrt{20}\, s_x \,;\, \bar{x} + \sqrt{20}\, s_x\right]$.

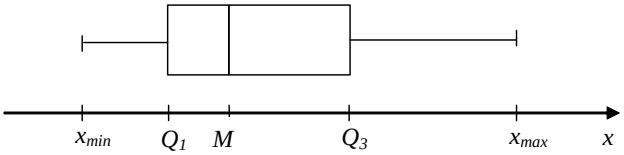
Remarquons aussi que seul le cas $t > 1$ est utile ; en effet, pour $0 < t \leq 1$, on a $1 - \frac{1}{t^2} \leq 0$. En particulier, on n'obtient pas de résultat intéressant pour $t = 1$, c'est-à-dire pour l'intervalle moyen $\left[\bar{x} - s_x \,;\, \bar{x} + s_x\right]$.

Vérifier les deux résultats ci-dessus avec les données de l'exercice 4.

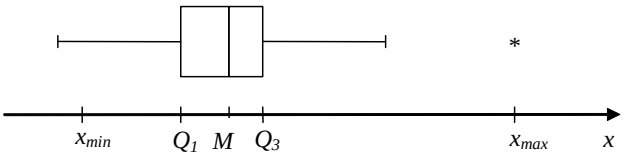
Exercice 6.

Pour une série statistique données, les trois quartiles, ainsi que les valeurs extrêmes de la série, peuvent être représentées graphiquement à l'aide de **boîtes à moustaches** (ou à dispersion, Box plots en anglais). Ce sont des représentations graphiques d'un caractère quantitatif résumé par ses valeurs extrêmes $x_{\min}$ et $x_{\max}$, et ses quartiles Q_1, $Q_2 = M$ et Q_3. Sur une échelle horizontale (ou verticale) :

- on trace un rectangle qui s'étend du premier au dernier quartile ;
- on partage ce rectangle par un segment tracé au niveau de la médiane ;
- on ajoute les "moustaches", c'est-à-dire des segments s'étendant de la valeur minimale au premier quartile, et du dernier quartile à la valeur maximale.



Une variante couramment utilisée consiste à prendre des moustaches de longueur $1,5 \times (Q_3 - Q_1)$; si la série comporte des valeurs extérieures aux moustaches, il s'agit de valeurs "aberrantes" au point de vue statistique, qu'on représente par exemple par *



Ces représentations permettent de comparer facilement différentes séries statistiques selon cinq paramètres de position (valeurs extrêmes et quartiles) et d'illustrer leur dispersion en mettant en évidence l'intervalle interquartile et l'étendue de chacune d'elles.

De plus, pour tenir compte de la taille de la série (nombre d'observations), on trace des rectangles d'une largeur proportionnelle à la racine carrée de celle-ci.

(Les 9 déciles et le 99 centiles partagent la série en 10 et 100 séries de même taille.)

Le tableau suivant donne la répartition des 100 techniciens supérieurs d'une grande entreprise de Picardie selon leurs salaires mensuels bruts (en euros) :

Salaires mensuels	Nombre de techniciens
$]1731;1962]$	13
$]1962;2193]$	35
$]2193;2424]$	21
$]2424;2655]$	13
$]2655;2886]$	9
$]2866;3117]$	5
$]3117;3348]$	3
$]3348;3579]$	1

1) Déterminer les valeurs extrêmes et les quartiles de cette série.

2) A l'aide de boîtes à moustaches, comparer cette entreprise aux trois entreprises, du même secteur industriel et de la même région, dont les caractéristiques sont donnés dans le tableau suivant :

Entreprise	Taille	$x_{\min}$	Q_1	Q_2	Q_3	$x_{\max}$
A	88	1770	2693	3385	3693	4539
B	81	1770	2308	2539	2924	4231
C	25	1462	1847	2077	2308	3462

Exercice 7.

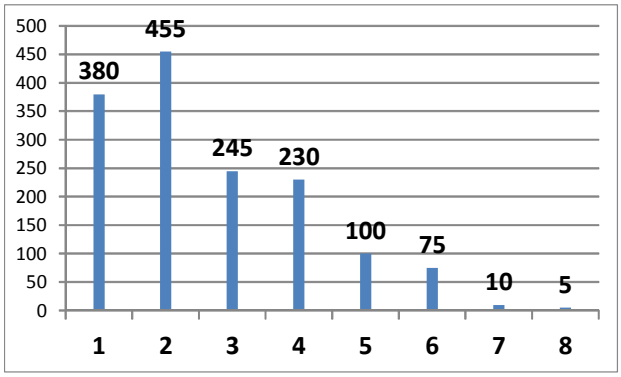
On désire comparer les distributions (groupées) des bénéfices nets hebdomadaires en euros de 2 magasins, sur 100 semaines comprenant toutes 6 jours d'ouverture.

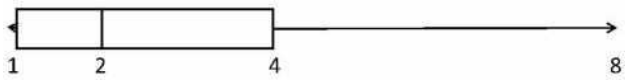
Magasin 1		Magasin 2	
Bénéfice	Nb de semaines	Bénéfice	Nb de semaines
1000	10	11000	8
2000	25	12000	24
3000	37	13000	38
4000	21	14000	20
5000	7	15000	10

1) Comparer les moyennes et les écart-types des deux distributions.
2) Calculer les coefficients de variation des deux ditributions et analyser les résultats.

Exercice 8.

Une enquête menée auprès de 1500 ménages d'une certaine région géographique rurale s'est intéressée à la variable correspondant à la taille du ménage, c'est-à-dire au nombre de personnes constituant le ménage. Les données recueillies peuvent être présentées sous la forme du diagramme en bâtons et de la boîte à moustaches ci-après.





On a par ailleurs déterminé que la taille moyenne des 1500 ménages est égale à 2,67, que la variance des tailles des ménages s'élève à environ 2,2678 et que le coefficient d'asymétrie de Fisher est égal à environ 0,829. Ces résultats sont-ils cohérents avec les diagrammes ci-dessus ?

Exercice 9.

1) A l'aide d'un tableur, on a effectué les calculs suivants sur deux séries statistiques :

x_i	n_i	$n_i x_i$	$n_i x_i^2$	$n_i((x_i-\bar{x})/s_x)^3$	$n_i((x_i-\bar{x})/s_x)^4$
1	100	100	100	-572,43	1024,00
2	150	300	600	-62,11	46,30
3	300	900	2700	7,95	2,37
4	150	600	2400	362,24	486,00
Total	**700**	**1900**	**5800**	**-264,35**	**1558,67**

C_i	x_i	n_i	$n_i x_i$	$n_i x_i^2$	$n_i((x_i-\bar{x})/s_x)^3$	$n_i((x_i-\bar{x})/s_x)^4$
[0 ; 2]	1	17	17	17	-32,99	41,16
]2 ; 5]	3,5	21	73,5	257,25	-21,58	21,78
]5 ; 10]	7,5	57	427,5	3206,25	-14,12	8,86
]10 ; 20]	15	55	825	12375	0,04	0,00
]20 ; 30]	25	35	875	21875	39,33	40,89
]30 ; 50]	40	15	600	24000	225,77	557,43
	Total	**200**	**2818**	**61730,5**	**196,44**	**670,13**

Pour chacune des deux séries, construire le diagramme des effectifs et déterminer les paramètres statistiques suivants : moyenne, écart-type, mode, médiane, quartiles, écart interquartile, coefficient de variation, coefficient d'asymétrie et coefficient d'aplatissement.

Vérifier la cohérence entre les diagrammmes et les paramètres obtenus.

2) Montrer que la série suivante est moins aplatie qu'une distribution normale, c'est-à-dire que $K' < 0$.

Revenus	]0,100]	]100,200]	]200,300]
n_i	3	5	2