LA STATISTIQUE DESCRIPTIVE

Module Bio statistiques
Niveau L_3

Biostatistique L_3

La statistique descriptive

1.1 Introduction.

La statistique est une branche des mathématiques qui s'intéresse à l'étude des phénomènes aléatoires, en contre opposé aux mathématiques déterministes plus familières aux étudiant(e)s, que sont le calcul différentiel, calcul intégral, géométrie, algèbre,…Le mot statistique dont le nom est dérivé de ''state'' en référence à tout ce qui est étatique, est relativement nouveau, puisqu'il a été introduit en Allemagne au XVIIème siècle. Par contre la pratique de la statistique est plus ancienne, elle fut utile aux grands empires en Mésopotamie, dans l'Égypte ancienne, ainsi que chez les romains et les empires indiens et chinois. Il s'agissait de bien connaître la population pour administrer sa répartition sur les territoires, collecter les impôts et gérer les aspects militaires.
De nos jours, on ne peut trouver un domaine qui peut être compris, analysé sans les méthodes statistiques. Que ce soit dans le domaine des sciences sociales, sciences de la vie ou sciences de l'ingénieur, les méthodes statistiques sont omniprésentes pour mettre de l'ordre dans le protocole de travail, elles permettent quand on est devant un chaos apparent des données, de déterminer par où commencer et quelles sont les étapes à suivre selon le contexte pour analyser ces données.

La statistique grosso-modo est formée de trois grandes classes : la statistique descriptive, la statistique inférentielle et la nouvelle branche qu'est la statistique exploratrice. Ce chapitre est consacré à la statistique descriptive. La statistique descriptive comme son nom l'indique, se propose de décrire les données, de les classer et de les présenter sous des formes claires et compréhensibles. Elle est à la base par exemple de toute organisation du système d'information d'une entreprise : statistiques de la production ou des ventes, statistiques financières, statistiques des ressources humaines…Elle est aussi une importante composante en sciences humaines de ce qu'on appelle les méthodes quantitatives. On va commencer par définir le lexique qu'on va utiliser tout le long de ce chapitre et même de ce livre.

1.2 : Les concepts et le vocabulaire de base.

Au début de tout travail statistique, il faut cerner avec précision sur quoi va porter l'étude. L'ensemble de tous les éléments sur lesquels porte l'étude s'appelle **population.** Une population peut être un ensemble d'êtres vivants (humains, oiseaux, poissons, bactéries,…) ou un ensemble de choses (maisons, voitures, rivières,…) ou un ensemble de faits (pannes, accidents, divorces,…). Chaque élément d'une population s'appelle **individu** ou **unité statistique**. Une population peut être finie (population d'un pays) ou presque infinie (population

des insectes), on considère généralement les populations comme finies mêmes si elles sont très grandes. Le nombre d'unités statistiques dans une population s'appelle **taille de la population** et on le note par **N**.
Quand une étude porte sur toute la population, on dit qu'on fait un **recensement**. Mais pour des raisons techniques ou économiques, il n'est généralement pas possible de collecter des données sur tous les éléments d'une population. Alors on se contente d'extraire une partie de la population appelée **échantillon** et restreindre l'étude à cet échantillon. On verra dans le chapitre V, qu'il existe des méthodes spécifiques permettant de s'assurer que l'échantillon soit représentatif de la population, c'est-à-dire une réplique en miniature de ce qui se passe dans la population. Pour l'instant, on suppose qu'on dispose d'un échantillon sur lequel porte l'étude (sans savoir comment il a été extrait). Le nombre d'éléments dans l'échantillon s'appelle **taille de l'échantillon** et sera noté par **n**.
On appelle **variable** tout caractère observé ou mesuré sur chacun des éléments de l'échantillon. On va réserver les dernières lettres de l'alphabet pour noter les variables : X, Y, Z, U...

Les différentes valeurs que prend une variable s'appellent **modalités**. Afin que le classement d'une unité statistique soit toujours possible sans ambiguïté, les différentes modalités doivent être à la fois incompatibles (un individu ne peut avoir plusieurs modalités à la fois) et exhaustives (tous les cas doivent être prévus). Il existe deux types de variables : Les **variables qualitatives** et les **variables quantitatives**. Une variable est dite qualitative si elle ne peut être mesurée ou quantifiée, mais peut être classée en catégories comme le sexe, la race, l'espèce, le niveau scolaire,…. Une variable est de type quantitatif si elle peut être mesurée ou quantifiée, comme le poids, la hauteur, le revenu, le nombre d'enfants, le nombre de pannes.
Les variables qualitatives sont constituées de deux sous-classes :

- Les variables qualitatives **nominales** : ce sont celles dont les modalités ne peuvent qu'être constatées, nommées.
  Exemple : Le sexe (masculin, féminin), la nationalité (Canadienne, Française, Marocaine,..), les cours suivis durant une session (mathématiques, anglais, philosophie,..) …
- Les variables qualitatives **ordinales**. ce sont les variables qualitatives dont les modalités appellent naturellement un ordre dans leur rangement.
  Exemple : Le niveau scolaire (primaire, secondaire, collégial, universitaire), le comportement lors d'une réception (incongru, correct, parfait,..), …

Les variables quantitatives sont elles aussi subdivisées en deux sous-classes :

- Les variables quantitatives **discrètes :** ce sont celles dont les modalités sont des valeurs isolées.
  Exemple : Le nombre de pannes, le nombre d'accidents, le nombre d'enfants,…

- Les variables quantitatives **continues**, ce sont celles dont les modalités forment un continuum. Ce sont celles qui peuvent prendre n'importe quelle valeur dans un intervalle raisonnable.
  Exemple : La taille, le poids, le revenu,…

1.3 Les échelles de mesures.

Pour les variables qualitatives, il existe deux échelles de mesure. **L'échelle nominale** qui s'adresse aux variables qualitatives nominales, elle ne sert qu'à coller une étiquette aux unités statistiques, elle ne les classe pas sur une échelle à une dimension.

Exemple 1.3.1 :

- X= sexe, alors X est une variable qualitative nominale et son échelle est nominale.
- Y=le numéro du dossard d'un joueur de hockey. Même si Y prend des valeurs numériques, ce n'est qu'une variable nominale et son échelle est nominale. Car on peut tout aussi bien mettre des lettres sur leur dossard ou des dessins.

L'autre **échelle est l'échelle ordinale** et s'adresse aux variables qualitatives ordinales, on l'appelle comme cela car il y a un ordre entre ses modalités.

Exemple 1.3.2 :

- X= le niveau scolaire d'une personne adulte, alors ses modalités peuvent être : primaire, secondaire, collégial, universitaire. Il y a un ordre chronologique entre ces modalités.
- Y= la note finale obtenue dans un cours de statistique, ses modalités seront : F, E, D, C, B, A ou A+. Il y a un ordre de mérite entre ces modalités.

Pour les variables quantitatives, il existe aussi deux types d'échelles, la première échelle est **l'échelle d'intervalle**. On l'appelle comme ça car la seule opération possible est la différence. On reconnaît une échelle d'intervalle par l'absence du zéro absolu (c'est-à-dire que si X=0, cela ne veut pas dire absence de ce qu'on mesure).

Exemple 1.3.3 :

- T= la température en degrés Celsius. Le jour où T=0°C, ça ne veut pas dire absence de température. Si on considère deux journées où la température est respectivement égale à 10 et 30 degrés, ça veut seulement dire qu'il y a un écart de 20 degrés entre ces deux journées. Si on prend deux sots d'eau où la température est respectivement égale à 35 et 45 degrés, si on les mélange, on ne va pas obtenir une eau chauffée à 80 degrés. Alors l'échelle de cette variable est une échelle d'intervalle.

- X=la date de naissance, si on est en 2010 et qu'on considère une personne née en 1950 et une autre née en 1980, tout ce qu'on peut dire est qu'il y a une différence d'âge de 30 ans entre elles. On ne peut pas dire que l'une est deux fois plus âgée que l'autre, car l'année prochaine ce ne serait plus vrai. Alors l'échelle de cette variable est une échelle d'intervalle.

L'autre échelle est **l'échelle de rapports**. C'est l'échelle la plus maniable, la plus riche. Elle admet un zéro absolu, c'est-à-dire si la variable est nulle, cela signifie l'absence de ce qu'on mesure. On peut faire toutes les opérations algébriques avec une telle échelle.

Exemple : 1.3.4 :

- X=le revenu familial annuel (en dollars), si X=0 cela veut dire qu'il n'y a pas eu de revenu. Si on prend deux familles dont le revenu respectif est de 30 000 et 120 000 dollars, on peut dire qu'il y a un écart de 90 000 dollars entre ces deux revenus, on peut aussi dire que la deuxième famille gagne 4 fois plus que la première. Si on additionne ces deux revenus, on aura un revenu global de 150 000 dollars. Alors l'échelle de cette variable est une échelle de rapports.
- Y=le nombre d'enfants dans un ménage. Si Y=0 cela veut dire que cette famille n'a pas d'enfant. On peut faire toutes les opérations algébriques avec les modalités de cette variable, donc son échelle est une échelle de rapports.

1.4 Les tableaux et graphiques.

Dans ce paragraphe on va détailler comment résumer l'information contenue dans une série de données soit par des tableaux ou des graphiques. On va commencer par les variables qualitatives.

1.4.1 Cas de variables qualitatives.

On va considérer deux exemples où on a des variables qualitatives observées sur un échantillon et suivre le traitement possible de ces données.

Exemple 1.4.1.1 : On a pris un échantillon de 50 achats de boissons non-alcoolisées achetées dans une grande surface, en notant par :

CC=Coca-Cola; S=Sprite; CL=Coke-Light; P=Perrier; PC=Pepsi-Cola. On a obtenu les résultats suivants.

CC S PC CL CC CC PC CL CC CL CC CC CC CL PC CC
CC P P S CC CL PC CL PC CC PC PC CC PC CC CC PC
P PC PC S CC CC CC S P CL P PC CC PC S CC CL

Alors ici la variable est X=Boisson non-alcoolisée, qui est une variable qualitative nominale. Pour présenter ces données sous forme de tableau, on dresse un tableau, dans la première colonne on énumère les cinq modalités de la

variable, dans la seconde colonne on donne la f**réquence absolue** ou l'effectif de chacune des modalités (c'est-à-dire le nombre de fois que cette modalité se répète dans l'échantillon) et dans la troisième colonne, on donne **la fréquence relative** de chacune des modalités. La **fréquence relative** d'une modalité étant égale à sa fréquence absolue divisée par la taille de l'échantillon. Ce qui donne :

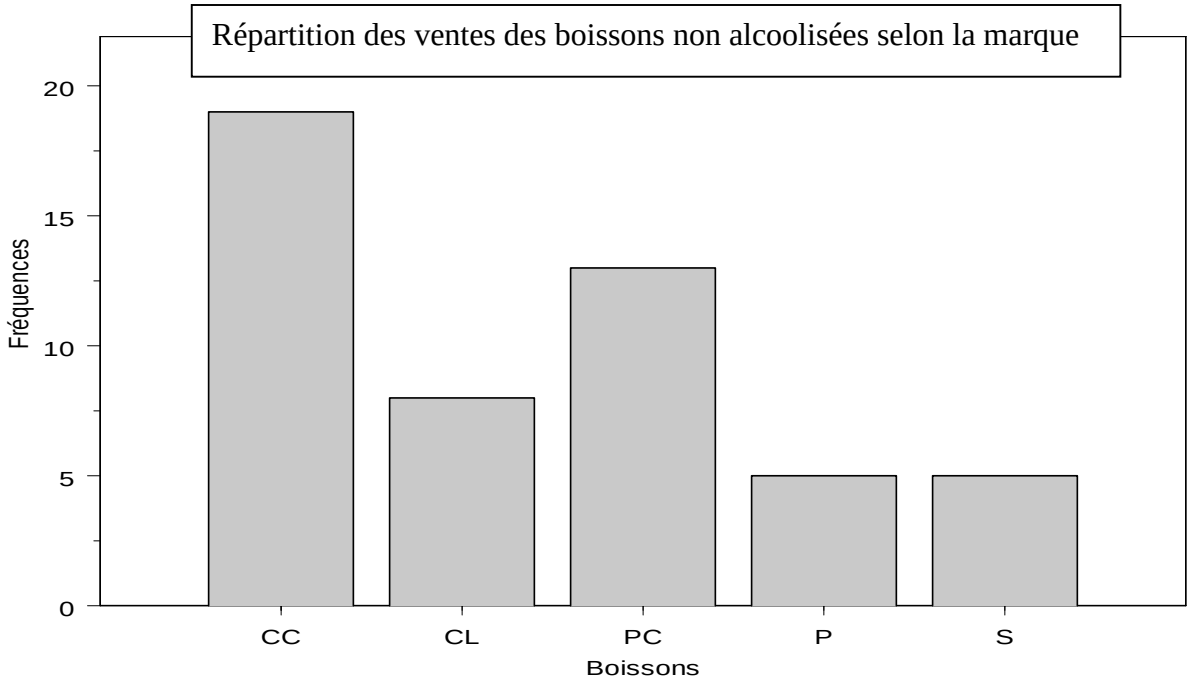
Tableau des fréquences des boissons non-alcoolisées		
X=Boisson	Fréquences absolues	Fréquences relatives
CC	19	0,38
CL	8	0,16
PC	13	0,26
P	5	0,10
S	5	0,10
Total	n=50	1
Source : données fictives.		

Ce tableau s'appelle **tableau de fréquences** de la variable.

Remarque : Pour une présentation complète des tableaux et graphiques, on doit mettre le titre en haut et la source des données en bas.

En ce qui concerne la représentation graphique, on va donner deux graphiques qui résument la même information contenue dans le tableau des fréquences.

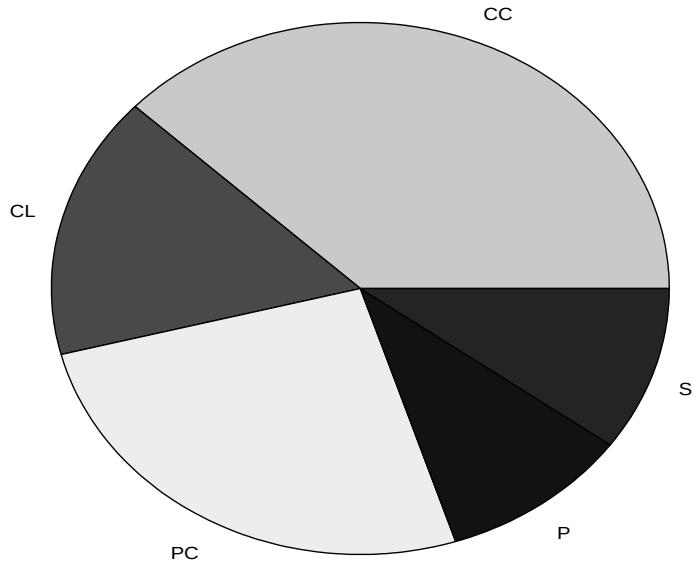
- Le diagramme à barres (horizontales ou verticales). Où on met sur un axe les modalités de la variable et sur l'autre axe les fréquences absolues ou les fréquences relatives.



Remarque : Les largeurs des barres doivent être les mêmes pour une belle esthétique du graphique, ainsi que la distance entre les bandes. On peut aussi ajouter les fréquences absolues au dessus des bandes.

- Le deuxième graphique qu'on peut faire est **le diagramme à secteurs** (ou circulaire) qui est une sorte de tarte où chaque modalité occupe une partie qui reflète sa fréquence relative.

Diagramme circulaire donnant la répartition des boissons non alcoolisées selon la marque



Exemple 1.4.1.2 : Lors d'une enquête de satisfaction de la clientèle, une compagnie de courtage a demandé à un échantillon de 60 clients d'indiquer leur degré de satisfaction vis-à-vis de leur conseiller financier, sur une échelle de 1 à 7, le 1 correspondant à <<pas du tout satisfait>> et le 7 correspondant à << extrêmement satisfait>>. On a obtenu les résultats suivants :

5 7 6 6 7 5 5 7 3 6 7 7 6 6 6 5 5 6 7 7
6 6 4 4 7 6 7 6 7 6 5 7 5 7 6 4 7 5 7 6
6 5 3 7 7 6 6 6 6 5 5 6 6 7 7 5 6 6 6 6

Ici la variable, ``degré de satisfaction`` est une variable qualitative ordinale. On peut résumer l'information contenue dans ces données sous forme d'un tableau de fréquences ce qui donne :

Tableau des fréquences du degré de satisfaction des clients.		
Degré de satisfaction	Fréquences absolues	Fréquences relatives
1	0	0,0000
2	0	0,0000
3	2	0,0333
4	3	0,0500
5	12	0,2000
6	25	0,4167
7	18	0,3000
Total	n=60	1,0000
Source : Données fictives.		

En ce qui concerne la représentation graphique, les mêmes graphiques qu'on a utilisés pour une variable qualitative nominale font l'affaire. Ce qui donne :

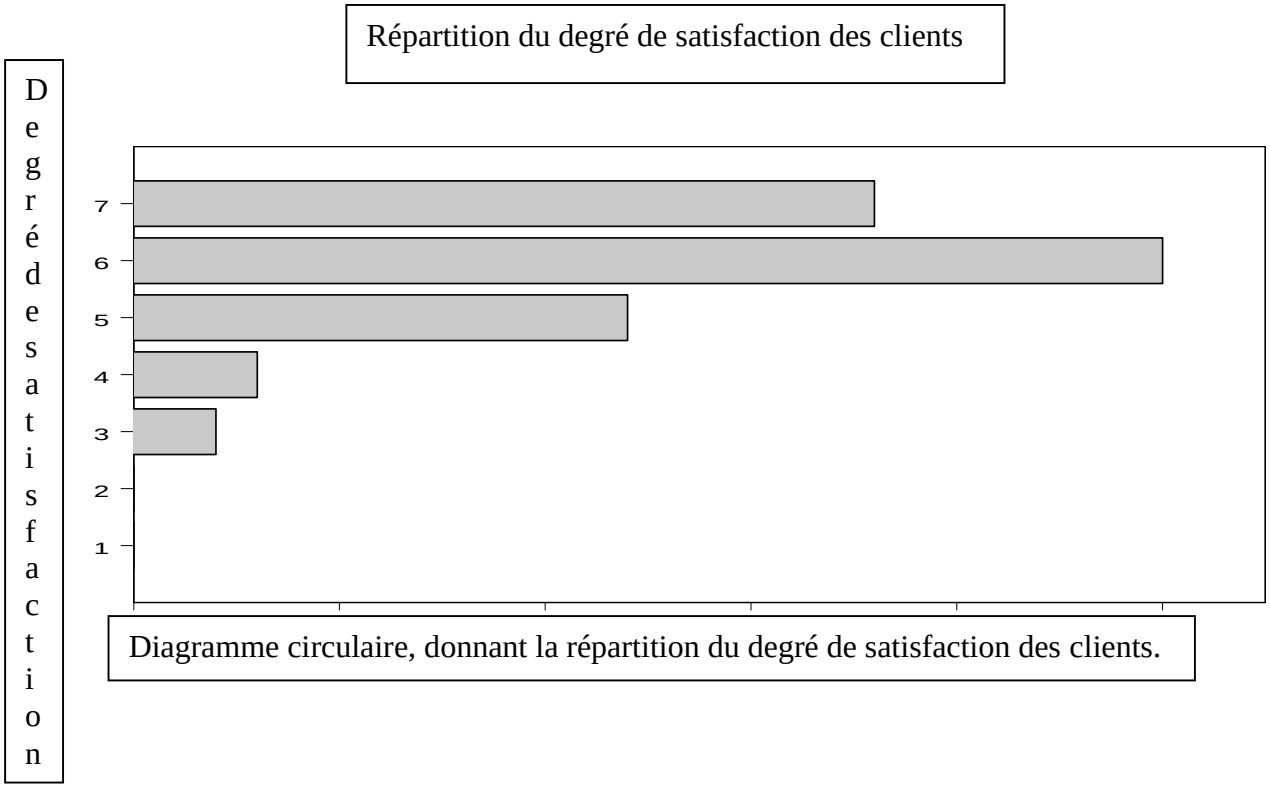


Diagramme circulaire, donnant la répartition du degré de satisfaction des clients.

1.4.2 Cas de variables quantitatives.

Le traitement des variables quantitatives discrètes étant différent de celui des variables quantitatives continues, on va donc réserver un sous paragraphe à chacune d'elles.

1.4.2.1 : Cas des variables quantitatives discrètes.

Soit X une variable quantitative discrète dont le nombre de modalités n'est pas trop grand. Alors on peut dresser un tableau des fréquences comme celui utilisé pour les variables qualitatives auquel on peut ajouter une colonne supplémentaire où on met les fréquences relatives cumulées au fur et à mesure qu'on ajoute une modalité de la variable. En ce qui concerne la représentation graphique, un seul graphique s'associe avec les variables quantitatives discrètes : **le diagramme à bâtons.**

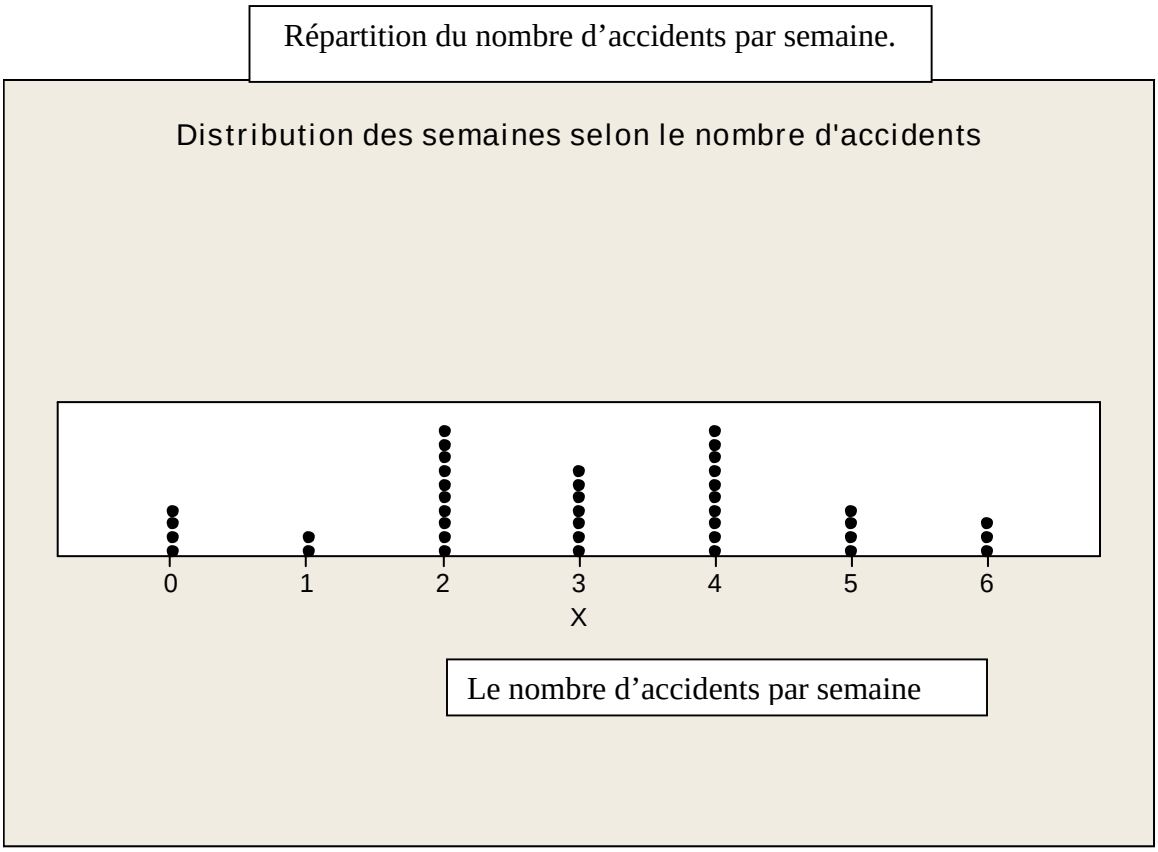
Exemple 1.4.2.1.1 : Un inspecteur en contrôle de qualité a extrait de sa base de données, un échantillon de 40 semaines où il a noté X, le nombre d'accidents de travail enregistrés par semaine. Il a obtenu les résultats suivants :

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

On peut donc dresser le tableau des fréquences suivant.

Tableau des fréquences du nombre d'accidents par semaine			
Le nombre d'accidents par semaine.	Fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
0	4	0,100	0,100
1	2	0,050	0,150
2	10	0,250	0,400
3	7	0,175	0,575
4	10	0,250	0,825
5	4	0,100	0,925
6	3	0,075	1,000
Total	n=40	1,000	

Quant au diagramme à bâtons, on obtient quelque chose comme :



Remarque : Les bâtons ne doivent pas avoir d'épaisseur, car la variable prend exactement les valeurs 0, 1, 2,…On peut ajouter les effectifs ou les fréquences relatives sur les bâtons.

1.4.2.2 : Cas de variables quantitatives continues.

Considérons maintenant un échantillon de données provenant d'une variable quantitative continue ou discrète avec un grand nombre de modalités. Il est donc inconcevable de dresser un tableau où on énumère les modalités d'une telle variable, il serait non analysable. Il faut donc grouper ces données en classes de valeurs. Deux questions se posent alors :

- Combien de classes faut-il former ?
- Quelles seront les largeurs de chacune des classes ?

La réponse à la première question, dépend de la taille de l'échantillon, le nombre de classe à former est donné par la formule de Sturges suivante :

$Le\ nombre\ de\ classes: K = 1 + \frac{10}{3} Log(n)$. Ainsi, par exemple, si n=150, il faut former $K = 1 + \frac{10}{3} Log(150) = 8{,}2536 \cong 9$ (on arrondit à l'entier immédiatement supérieur). Une fois qu'on sait combien de classes à former. On essaie de former des classes de même amplitude (largeur) et cette amplitude sera égale à

$$A = \frac{la\ plus\ grande\ observation - la\ plus\ petite\ observation}{K} = \frac{X_{max} - X_{min}}{K}.$$

On arrondit cette amplitude selon les données pour avoir des bornes de classes faciles à manipuler.

Exemple 1.4.2.2.1 : Soit X, les recettes quotidiennes(en dollars) d'un petit magasin. On a sélectionné un échantillon de taille n=40 jours au hasard qui ont donné les résultats suivants :

16,00 58,50 68,20 78,00 79,45 142,20 145,3 186,70 209,05 216,75
219,70 247,75 249,10 256,00 257,15 262,35 268,60 269,60 270,15 284,45
319,00 332,00 343,29 350,75 354,90 372,60 383,20 389,20 404,55 420,20
428,50 432,40 444,60 446,80 456,10 458,10 493,95 511,95 521,05 621,35

Le nombre de classe à former est $K = 1 + \frac{10}{3}\log(40) = 6{,}34 \cong 7\ classes$ d'amplitude chacune égale à $A = \frac{621{,}35-16}{7} = 86{,}48 \cong 90$. Cette amplitude est arrondie à 90. Ce qui donne le tableau des fréquences suivant, où les classes sont des intervalles fermés à gauche et ouverts à droite sauf le dernier qui est un intervalle fermé des deux côtés.

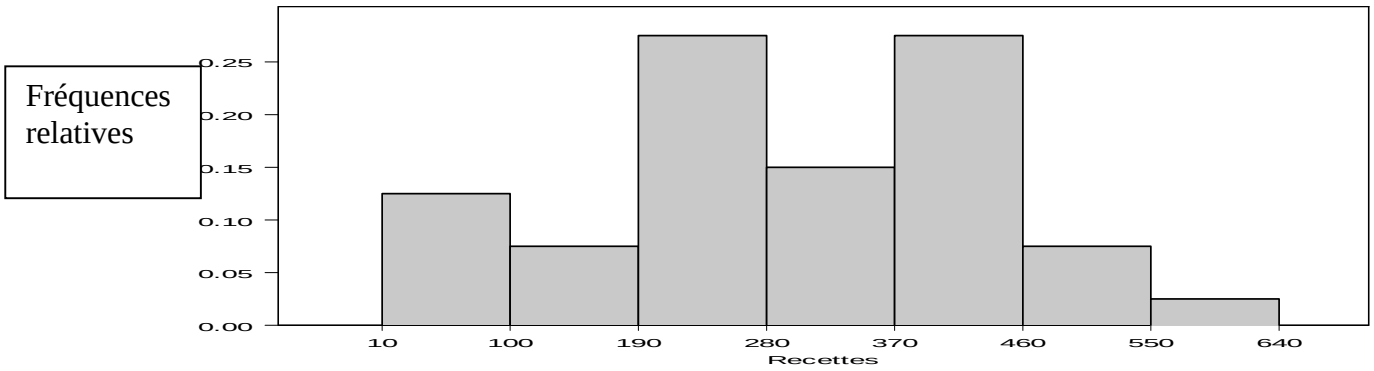
Répartition des 40 semaines selon les recettes hebdomadaires du dépanneur

X=les recettes	Fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
[10 ; 100[	5	0,125	0,125
[100 ;190[	3	0,075	0,200
[190 ;280[	11	0,275	0,475
[280 ;370[	6	0,150	0,625
[370 ;460[	11	0,275	0,900
[460 ;550[	3	0,075	0,975
[550 ;640]	1	0,025	1,000
Total	n=40	1,000	

Quand aux graphiques, on va ici préviligier trois graphiques pour les variables quantitatives continues.

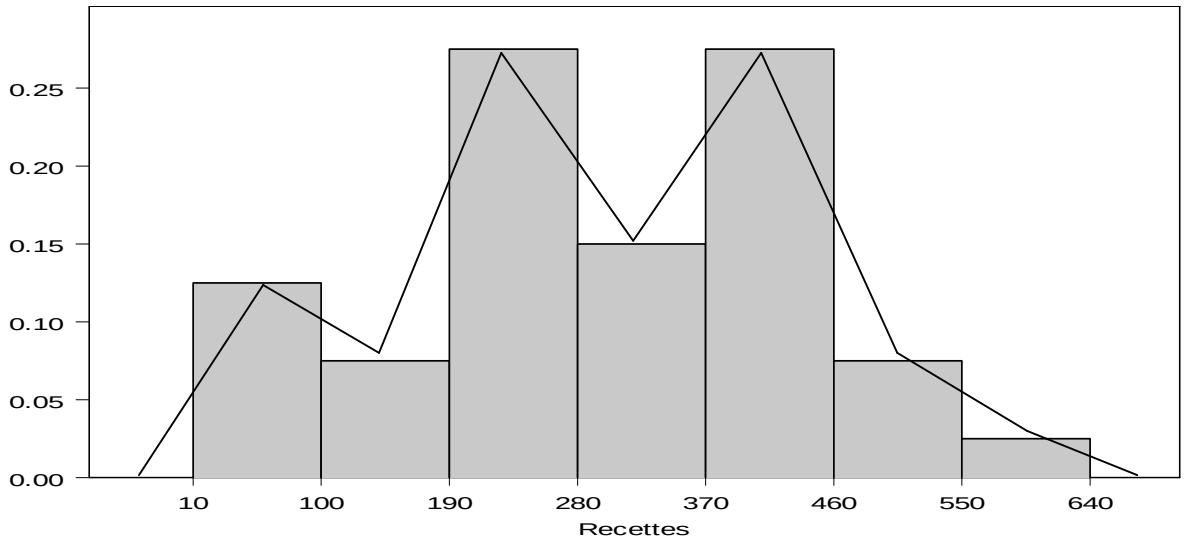
- **L'histogramme**, qui est une suite de rectangles juxtaposés les uns aux autres dressés au-dessus de chacune des classes, dont la largeur est égale à l'amplitude de la classe (prise comme unité de mesure) et dont la surface reflète la fréquence relative de la classe qu'il représente.

Histogramme donnant la répartition des 40 semaines en fonction des recettes hebdomadaires



- **Le polygone des fréquences**, qui consiste à joindre le milieux des sommets des rectangles d'un histogramme par une ligne en zig-zag et cette ligne se ferme en ajoutant aux deux extrémités deux classes fictives de même amplitude que les autres, comme ça la surface délimitée par l'histogramme est identique à celle délimitée par le polygone des fréquences. Le polygone de fréquences est très utile quand on veut comparer le comportement de la même variable mesurée sur plusieurs groupes (on peut penser à comparer le revenu des hommes et des femmes) ou la même variable mesurée sur le même échantillon à différents instants (on peut comparer le poids du même groupe à différents moments d'une diète).

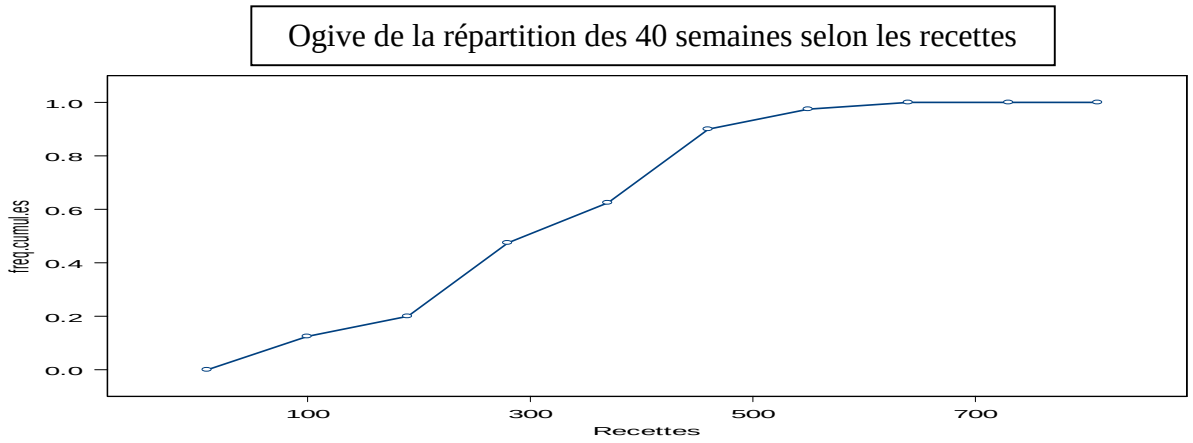
Polygone des fréquences donnant la répartition des 40 semaines selon les recettes hebdomadaires.



- **La courbe des fréquences cumulées (Ogive).**

Comme son nom l'indique, elle consiste à tracer le graphique des fréquences cumulées, en mettant les limites des classes sur l'axe horizontal et les fréquences

cumulées sur l'axe vertical, ces dernières se cumulant à la fin de chacune des classes. Ce graphique aura l'allure d'une courbe croissante variant entre 0 et 1.



Remarque : Lorsque les classes ne sont pas de même amplitude, il faut se rappeler que la surface du rectangle d'un histogramme étant égale à sa fréquence relative à la classe associée à ce rectangle, alors si la largeur de cette classe par exemple est le double de la l'amplitude de base, la hauteur du rectangle doit être divisée par deux.

1.5 : Les mesures de tendance centrale

On appelle mesures de tendance centrale, des valeurs de la variable susceptibles de nous donner une idée sur la donnée qui occupe le centre d'une série statistique. On va décrire dans ce paragraphe, les trois plus importantes mesures de tendance centrale que sont **le mode, la moyenne et la médiane**.

1.5.1.1 : Le mode

On appelle le mode d'une variable X, la valeur de la variable qui a la plus grande fréquence et on le note Mo(X). Le mode est une importante mesure de tendance centrale pour les variables qualitatives nominales.

Remarque : Une distribution peut avoir un seul mode et on dit qu'elle est unimodale, ou plusieurs modes et on dit qu'elle est multimodale.

Exemple 1.5.1.1.1 : Si on reprend l'exemple des boissons non-alcoolisées, on avait le tableau des fréquences suivant :

Tableau des fréquences des boissons non-alcoolisées		
X=Boisson	Fréquences absolues	Fréquences relatives
CC	19	0,38
CL	8	0,16
PC	13	0,26
P	5	0,10
S	5	0,10
Total	n=50	1

Alors, le mode de cette variable est Mo(X)=Coca-Cola (CC), cela signifie que dans cet échantillon, la boisson la plus fréquemment achetée est Coca-Cola.

Exemple 1.5.1.1.2 : En reprenant l'exemple des recettes quotidiennes d'un petit magasin, où la variable est quantitative continue avec des données groupées en classes, on avait le tableau des fréquences suivant :

X=les recettes	Fréquences absolues	Fréquences relatives
[10 ; 100[	5	0,125
[100 ;190[	3	0,075
[190 ;280[	11	0,275
[280 ;370[	6	0,150
[370 ;460[	11	0,275
[460 ;550[	3	0,075
[550 ;640]	1	0,025
Total	n=40	1,000

Ici, on voit qu'il y a deux classes qui ont les plus hautes fréquences, on les appelle des classes modales. Alors on est en présence d'une distribution de données bimodale, et les deux modes sont les milieux des deux classes modales, à savoir Mo(X)=235 et Mo(X)=415. Cela veut dire que dans cet échantillon les recettes quotodiennes les plus fréquentes sont soit de 235$ ou de 415$. Il y a des auteurs qui font des interpolations à l'intérieur des classes modales pour trouver le mode, on estime que c'est un effort inutile, vue que dans le cas d'une variable quantitative le mode joue un rôle très marginal. On voit que le mode d'une variable est une mesure de tendance centrale facile à déterminer et s'applique à tous les types de variables, mais sa portée comme mesure d'analyse est très limitée.

1.5.2 : La moyenne.

La moyenne arithmétique ou simplement la moyenne est la mesure de tendance centrale la plus connue. Elle ne s'applique qu'aux variables quantitatives. On va décrire la méthode pour calculer la moyenne d'une variable quantitative selon que les données sont en vrac, groupées par valeurs ou groupées par classes.

1.5.2.1 : Les données en vrac.

Soit X une variable quantitative dont les valeurs observées sur un échantillon forment une série en vrac $x_1, x_2, \ldots, x_n$ alors la moyenne de cet échantillon est

$$\bar{x} = \frac{x_1 + x_2 + \cdots . + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Exemple 1.5.2.1.1 : Un commerçant a l'habitude de noter dans son registre le nombre de clients qui se présentent quotidiennement à son magasin. On a pris un échantillon de taille 10 de ce registre et on trouvé les valeurs suivantes :

120 105 90 201 196 65 88 163 103 116

Alors dans cet échantillon le nombre moyen des clients qui se présentent à ce magasin par jour est donné par la formule suivante :

$$\bar{x} = \frac{x_1+x_2+\cdots.+x_n}{n} = \frac{120+105+\cdots.+116}{10} = 124{,}7 \text{ clients par jour.}$$

1.5.2.2 : Les données groupées par valeurs.

Soit X une variable quantitative discrète dont les données se présentent sous forme d'un tableau où elles sont classées par valeurs, supposons que la taille de l'échantillon est n et qu'il y a k valeurs différentes pour cette variable. Alors la moyenne d'un tel échantillon de données est :

$$\bar{x} = \frac{\sum(valeur) * (sa\ fréquence\ absolue)}{taille\ de\ l'échantillon} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

Exemple 1.5.2.2.1 : Reprenons les données de l'exemple 1.4.2.1.1, où X est le nombre d'accidents de travail par semaine. On avait le tableau de données suivant :

Tableau des fréquences du nombre d'accidents par semaine	
X	Fréquences absolues
0	4
1	2
2	10
3	7
4	10
5	4
6	3
Total	n=40

Alors la moyenne de cet échantillon est égale à

$$\bar{x} = \frac{(0)*(4)+(1)*(2)+\cdots..+(6)(3)}{40} = 3{,}025 \text{ accidents par semaine.}$$

1.5.2.3 : Les données groupées par classes.

Supposons qu'on est devant un tableau où les données provenant d'un échantillon sont groupées par classes. Alors pour calculer la moyenne de cet échantillon, on va utiliser une formule approximative, où chaque classe est assimilée à son centre et on utilise la même formule que pour le cas où les données sont groupées par valeurs. Si on note par m_i le milieu de la ième classe

et qu'on suppose que la taille de l'échantillon est n et qu'il y a k classes, alors la moyenne de l'échantillon est :

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$$

Exemple 1.5.2.3.1 : En reprenant l'exemple 1.4.2.2.1 où X est la recette quotidienne d'un petit magasin, on avait le tableau suivant auquel on a ajouté une colonne à gauche contenant le milieu des classes :

m_i	X=les recettes	Fréquences absolues
55	[10 ; 100[	5
145	[100 ;190[	3
235	[190 ;280[	11
325	[280 ;370[	6
415	[370 ;460[	11
505	[460 ;550[	3
595	[550 ;640]	1
	Total	n=40

Alors la moyenne de cet échantillon est :

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n} = \frac{(55) * (5) + \cdots + (595) * (1)}{40} = 298\ dollars.$$

1.5.2.4 : Les propriétés d'une moyenne échantillonnale.

Soit X une variable quantitative dont la moyenne échantillonnale est $\bar{x}$ et soit Y une autre variable quantitative transformée linéaire de X, c'est-à-dire que $Y = a + b * X$ où a et b sont des constantes réelles. Alors la moyenne échantillonnale de Y sera égale à

$\bar{y} = a + b * \bar{x}$.

On dit que la moyenne conserve la transformation linéaire entre les variables.

Exemple 1.5.2.4.1 : Soit X, le nombre d'heures qu'un étudiant travaille à temps partiel par semaine. Supposons qu'à partir d'un échantillon d'étudiants, on a pu trouver qu'en moyenne le nombre d'heures travaillées par ces étudiants est égale à $\bar{x} = 14{,}5$ heures/semaine. Si le salaire horaire est de 10$ et que les patrons de ces étudiants leur offrent 30$ par semaine pour leurs déplacements, quel est le gain net moyen hebdomadaire de ces étudiants ? Posons Y, le gain net hebdomadaire de ces étudiants alors $Y = 30 + 10 * X$, donc le gain moyen hebdomadaire de cet échantillon d'étudiants est égal à

$\bar{y} = 30 + 10 * \bar{x} = 30 + 10 * 14{,}5 = 175\$$.

1.5.3 : La médiane.

La médiane est la valeur de la variable qui divise l'échantillon en deux groupes d'égal effectif. Il y a 50% des données qui sont inférieures ou égales à la médiane et 50% des données qui sont supérieures ou égales à la médiane. La médiane se calcule pour des variables qualitatives ordinales et pour des variables quantitatives. On note la médiane d'une variable X par Med(X) ou par $\tilde{x}$. Dans ce qui suit on va décrire les façons de calculer une médiane dans les différents cas possibles.

1.5.3.1 : Cas d'une variable qualitative ordinale.

Puisque les modalités d'une telle variable sont déjà ordonnées par nature, alors pour déterminer la médiane, on calcule $l = (50\%) * n$, et donc

$$Med(X) = \tilde{x} = \begin{cases} \dfrac{x_{(l)} + x_{(l+1)}}{2} & si \quad l \ \ est\ un\ entier \\ x_{[l]+1} & si \quad l \ \ n'est\ pas\ un\ entier \end{cases}$$

Où $x_{[l]+1}$ signifie, l'observation occupant le rang immédiatement supérieur à l.

Exemple 1.5.3.1.1 : Reprenons les données de l'exemple 1.4.1.2, où X est le degré de satisfaction de la clientèle, on avait le tableau suivant :

Tableau des fréquences du degré de satisfaction des clients.	
Degré de satisfaction	Fréquences absolues
1	0
2	0
3	2
4	3
5	12
6	25
7	18
Total	n=60

Ici, n=60 et $l = (50\%) * n = 30 \ \ est\ un\ entier$, alors

$Med(X) = \tilde{x} = \frac{x_{(30)}+x_{(31)}}{2} = \frac{6+6}{2} = 6$. Le degré de satisfaction médian de la clientèle est égal à 6. Ce qui veut dire que dans cet échantillon 50% des clients ont un degré de satisfaction de 6 ou moins et l'autre 50% un degré de satisfaction de 6 ou plus.

1.5.3.2 : Cas de données quantitatives en vrac ou groupées par valeurs.

On doit d'abord ordonner les données par ordre croissant avant d'appliquer la même procédure que pour les variables qualitatives ordinales. Ci-après nous donnerons un exemple pour chacun de ces deux cas.

Exemple 1.5.3.2.1 : Reprenons les données de l'exemple 1.5.2.1.1 où la variable est le nombre de clients qui se présentent quotidiennement au magasin. On avait des données en vrac :

120 105 90 201 196 65 88 163 103 116

En les ordonnant, on aura : 65 88 90 103 105 116 120 163 196 201.

Ici, n=10 et $l = (50\%) * n = 5 \ est\ un\ entier$, alors
$Med(X) = \tilde{x} = \frac{x_{(5)}+x_{(6)}}{2} = \frac{105+116}{2} = 110{,}5$. Ce qui veut dire qu'à partir de cet échantillon, on peut affirmer que dans 50% des journées, ce magasin reçoit 110 clients ou moins par jour et dans l'autre 50% des journées, il reçoit 110 clients ou plus.

Exemple 1.5.3.2.2 : : Reprenons les données de l'exemple 1.4.2.1.1, où X est le nombre d'accidents de travail par semaine. On avait le tableau de données où les modalités de la variable sont groupées par valeurs, qu'on va changer un peu en ajoutant une donnée supplémentaire :

Tableau des fréquences du nombre d'accidents par semaine	
Nombre d'accidents par semaine	Fréquences absolues
0	4
1	2
2	10
3	7
4	10
5	4
6	4
Total	n=41

Ici, n=41 et $l = (50\%) * n = 20{,}5 \ n'est\ pas\ un\ entier$, alors
$Med(X) = \tilde{x} = x_{[20,5]+1} = x_{(21)} = l'observation\ qui\ occupe\ la\ 21ème\ position = 3.$
C'est-à-dire que dans cet échantillon, dans 50% des semaines, on observe 3 accidents ou moins par semaine et l'autre 50% des semaines, on observe 3 accidents ou plus par semaine.

1.5.3.3 : Cas de données groupées par classes.
Dans le cas où on dispose d'un tableau de fréquences complet (incluant les fréquences cumulées) des données groupées par classes. Il faut d'abord déterminer la classe médiane, qui est la classe où les fréquences cumulées dépassent pour la première fois 50%. Cette classe aura la forme :

$C_m = [b_{inf}; b_{sup}[$, alors on obtient la médiane par interpolation à l'intérieur de cette classe médiane et on obtient la formule suivante :

$$Med(X) = \tilde{x} = b_{inf} + \frac{(0,5 - F_{(m-1)})}{f_{r,m}} * A_m \text{ où}$$

$b_{inf} = borne\ inférieure\ de\ la\ classe\ médiane.$
$F_{(m-1)} = la\ fréquence\ cumulée\ avant\ la\ classe\ médiane.$
$f_{r,m} = la\ fréquence\ relative\ de\ la\ classe\ médiane.$
$A_m = l'amplitude\ de\ la\ classe\ médiane.$

Exemple 1.5.3.3.1 : En reprenant les données où X donne la recette quodienne d'un petit magasin, on retrouve le tableau des fréquences suivant :

X=les recettes	Fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
[10 ; 100[	5	0,125	0,125
[100 ;190[	3	0,075	0,200
[190 ;280[	11	0,275	0,475
[280 ;370[	6	0,150	0,625
[370 ;460[	11	0,275	0,900
[460 ;550[	3	0,075	0,975
[550 ;640]	1	0,025	1,000
Total	n=40	1,000	

Alors ici, la classe médiane est $C_m = [b_{inf}; b_{sup}[$=[280 ;370[
$b_{inf} = 280 \quad F_{(m-1)} = 0,475$
$f_{r,m} = 0,150 \quad A_m = 90$ ce qui donne une médiane égale à :

$$\tilde{x} = b_{inf} + \frac{(0,5 - F_{(m-1)})}{f_{r,m}} * A_m = 280 + \frac{(0,5 - 0,475)}{0,150} * 90 = 295\$$$

Ce qui veut dire qu'en se basant sur cet échantillon de données, 50% des recettes quotidiennes de ce petit magasin sont inférieures ou égales à 295$ et les autres 50% sont supérieures ou égales à 295$.

Remarque 1 : Le calcul de la médiane est basé sur l'ordre des observations et non sur leur valeur. Contrairement à la moyenne, la médiane est insensible aux données extrêmes. Dans le cas où les données sont très différentes, la médiane est une meilleure mesure de tendance centrale.

Remarque 2 : Si pour une variable X quantitative les 3 mesures de tendance centrale sont presque égales, on dit alors que la variable est symétrique et alors n'importe laquelle de ces mesures peut être utilisée comme mesure de cette

tendance centrale. S'il y a un grand écart entre ces mesures alors c'est la médiane qu'on doit priviligier.

1.6 : Les mesures de position.
On a déjà parlé de la médiane comme mesure de tendance centrale, mais elle est aussi une mesure de position car elle permet de diviser une série d'observations en deux groupes chacun contenant 50% de données. On va définir d'autres mesures de position qui permettent d'autres découpages d'une série d'observations.

1.6.1 : Les quartiles. Lorsqu'on veut diviser les données en quatres groupes, chacun contenant 25% des observations, on utilise des mesures appelées quartiles.
Q_1=le 1er quartile, à sa gauche il y a 25% des observations, qu'on note Q_1.
Q_2=le 2ème quartile, coincide avec la médiane, qu'on note $Q_2 = Med(X)$.
Q_3=le 3ème quartile, à sa gauche il y a 75% des observations, qu'on note Q_3.
On va décrire la façon de les calculer, dans les 3 cas possibles pour une variable quantitative.

1.6.1.1 : Les données en vrac. On suit les étapes suivantes.
Étape 1 : On ordonne les données par ordre croissant.
Étape 2 : On calcule l'indice $l = (i\%) * (n)$ où i est le pourcentage correspondant à la mesure voulue et n est le nombre d'observations.
Étape 3 : (a) si l n'est pas un entier, alors le ième quartile est égal à l'observation occupant la position immédiatement supérieure à l.
(b) si l est un entier, alors le ième quartile est la moyenne des observations occupant les positions l et $(l+1)$.

Exemple 1.6.1.1.1 : n=12 et les observations sont :
-2 -3 10 12 120 11 4 8 6 13 130 200.
Étape 1 : -3 -2 4 6 8 10 11 12 13 120 130 200.
Étape 2 : Si on veut déterminer Q_1, on calcule $l_1 = (25\%) * (n) = 3$.
Si on veut déterminer Q_2, on calcule $l_2 = (50\%) * (n) = 6$.
Si on veut déterminer Q_3, on calcule $l_3 = (75\%) * (n) = 9$.
Étape 3 : Puisque l_1 est un entier alors $Q_1 = \frac{la\ 3ème\ obs + la\ 4ème\ obs}{2} = \frac{4+6}{2} = 5$.
Puisque l_2 est un entier alors $Q_2 = \frac{la\ 6ème\ obs + la\ 7ème\ obs}{2} = \frac{10+11}{2} = 10{,}5$.
Puisque l_3 est un entier alors $Q_3 = \frac{la\ 9ème\ obs + la\ 10ème\ obs}{2} = \frac{13+120}{2} = 66{,}5$.

Exemple 1.6.1.1.2 : n=10 et les observations sont :
3 10 12 8 6 100 15 6 3 14.
Étape 1 : 3 3 6 6 8 10 12 14 15 100

Étape 2 : Si on veut déterminer Q_1, on calcule $l_1 = (25\%) * (n) = 2{,}5$.
Si on veut déterminer Q_2, on calcule $l_2 = (50\%) * (n) = 5$.
Si on veut déterminer Q_3, on calcule $l_3 = (75\%) * (n) = 7{,}5$.
Étape 3 : Puisque l_1 n'est pas un entier alors $Q_1 = la\ 3ème\ observation = 6$.
Puisque l_2 est un entier alors $Q_2 = \frac{la\ 5ème\ obs + la\ 6ème\ obs}{2} = \frac{8+10}{2} = 9$.
Puisque l_3 n'est pas un entier alors $Q_3 = la\ 8ème\ observation = 14$.
Remarque : La procédure décrite pour trouver les quartiles est une convention parmi d'autres. Il n'y a pas d'accord général sur la méthode à utiliser pour déterminer les quartiles. Si vous utilisez des logiciels, les valeurs trouvées diffèrent d'un logiciel à l'autre. Par exemple, si on prend la série en vrac suivantes : 1 3 6 10 15 21 28 36, alors la calculatrice TI-83 et plus et les logiciels suivants donnent :

logiciel	Q_1	Q_2	Q_3
SPSS	3,75	12,5	26,25
SAS	4,5	12,5	24,5
STATDISK	4,5	12,5	24,5
Excel	5,25	12,5	22,75
R	5,25	12,5	22,75
Splus	5,25	12,5	22,75
Minitab	3,75	12,5	26,25
TI-83 et plus	4,5	12,5	24,5

Heureusement, dans la pratique, les échantillons sont très grands et ces fluctuations ne changent pas grand-chose dans les analyses des données.

1.6.1.2 : Les données groupées par valeurs.

On suit la même démarche que dans le cas des données en vrac, sauf l'étape 1 qui devient inutile, puisque les données sont en général déjà ordonnées par ordre croissant.

Exemple 1.6.1.2.1 : En reprenant le tableau de l'exemple 1.5.3.2.2, déterminer les 3 quartiles de la variable X=le nombre d'accidents par semaine.

Tableau des fréquences du nombre d'accidents par semaine	
X	Fréquences absolues
0	4
1	2
2	10
3	7
4	10
5	4
6	4
Total	n=41

Réponse :
Étape 2 : Si on veut déterminer Q_1, on calcule $l_1 = (25\%) * (n) = 10{,}25$.
Si on veut déterminer Q_2, on calcule $l_2 = (50\%) * (n) = 20{,}5$.
Si on veut déterminer Q_3, on calcule $l_3 = (75\%) * (n) = 30{,}75$.
Étape 3 : Puisque l_1 n'est pas un entier alors $Q_1 = la\ 11ème\ observation = 2$.
Puisque l_2 n'est pas un entier alors $Q_2 = la\ 21ème\ observation = 3$.
Puisque l_3 n'est pas un entier alors $Q_3 = la\ 31ème\ observation = 4$.
$Q_1 = 2$ signifie que dans cet échantillon, durant 25% des semaines, on a observé 2 accidents par semaine ou moins.
$Q_2 = 3$ signifie que dans cet échantillon, durant 50% des semaines, on a observé 3 accidents par semaine ou moins.
$Q_3 = 4$ signifie que dans cet échantillon, durant 75% des semaines, on a observé 4 accidents par semaine ou moins.

1.6.1.3 : Les données groupées par classes.
On suit la même démarche utilisée pour calculer la médiane quand les données sont groupées par classes. On détermine la classe où on a dépassé le pourcentage relatif à chaque quartile et on fait une interpolation à l'intérieur de cette classe. On aboutit à la même formule que celle de la médiane où seul le poucentage est à adapter.

Exemple 1.6.1.3.1 : En reprenant les données de l'exemple 1.5.3.3.1, déterminer les 3 quartiles de la variable X, soit les recettes quotidiennes d'un petit dépanneur, et interpréter ces mesures.

X=les recettes	Fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
[10 ; 100[	5	0,125	0,125
[100 ;190[	3	0,075	0,200
[190 ;280[	11	0,275	0,475
[280 ;370[	6	0,150	0,625
[370 ;460[	11	0,275	0,900
[460 ;550[	3	0,075	0,975
[550 ;640]	1	0,025	1,000
Total	n=40	1,000	

Réponse :
(a) Pour déterminer le premier quartile, les fréquences relatives cumulées ont dépassé 25% pour la première fois au niveau de la classe [190 ; 280[, donc $Q_1 = 190 + \frac{(0{,}25-0{,}20)}{0{,}275} * 90 = 206{,}36\$$. Ce qui signifie que dans cet échantillon de données, 25% des journées, les recettes quotidiennes de ce petit magasin ont été de 206,36$ ou moins.

(b) Pour déterminer le deuxième quartile (on refait ce qu'on a déjà fait pour calculer la médiane), les fréquences relatives cumulées ont dépassé 50% pour la première fois au niveau de la classe [280 ; 370[, donc $Q_2 = 280 + \frac{(0,50-0,475)}{0,150} * 90 = 295\$$. Ce qui signifie que dans cet échantillon de données, 50% des journées, les recettes quotidiennes de ce petit magasin ont été de 295$ ou moins.

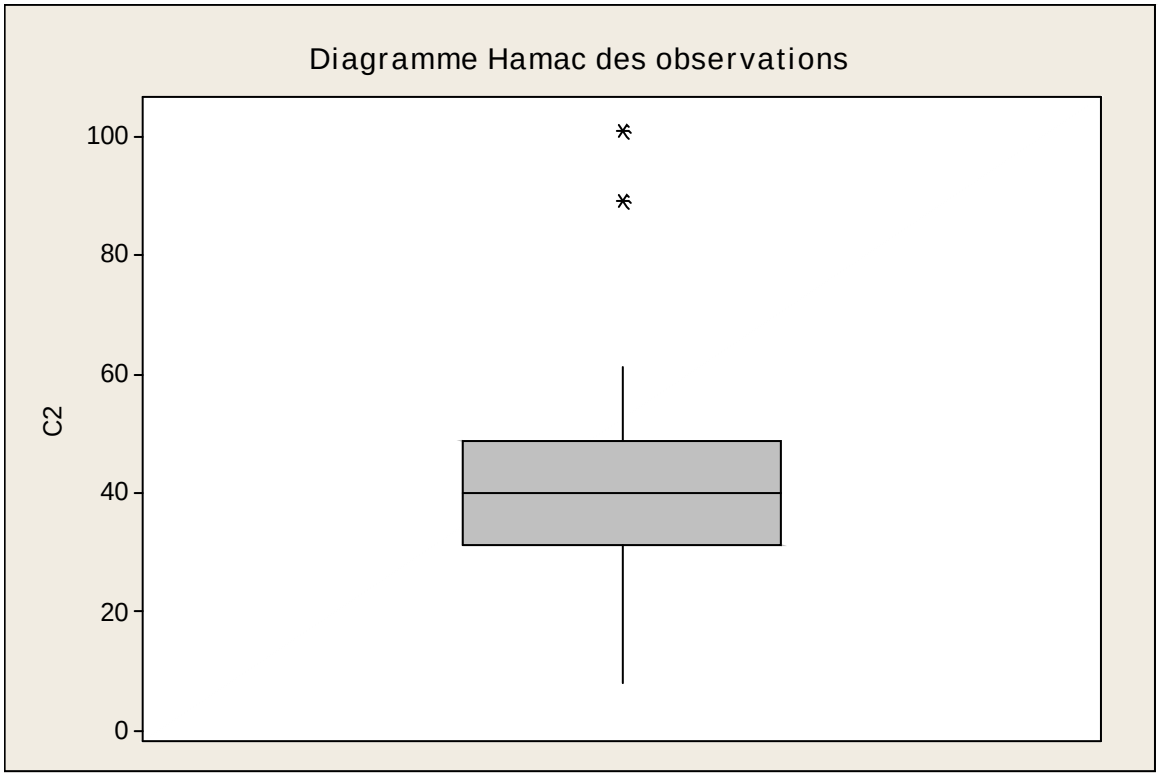
(c) Pour déterminer le troisième quartile, les fréquences relatives cumulées ont dépassé 75% pour la première fois au niveau de la classe [370 ; 460[, donc $Q_3 = 370 + \frac{(0,75-0,625)}{0,275} * 90 = 410,91\$$. Ce qui signifie que dans cet échantillon de données, 75% des journées, les recettes quotidiennes de ce petit magasin ont été de 410,91$ ou moins.

Utilité des quartiles. Les quartiles, en plus de leur utilisation comme mesures de position, s'utilisent pour détecter des données aberrantes dans toute série de données. Cette détection se fait à l'aide d'un graphique, appelé graphique en boîte (box-plot) ou hamac ou diagramme à moustache selon les auteurs. Son principe consiste à calculer les quartiles de la série et deux limites acceptables. Soient une limite inférieure $L_{inf} = Q_1 - 1,5 * (Q_3 - Q_1)$ et une limite supérieure $L_{sup} = Q_3 + 1,5 * (Q_3 - Q_1)$. Toute observation qui ne se trouve pas entre ces deux limites est jugée aberrante et doit être exclue de la série avant toute analyse des données (on essaye de faire une interprétation de la présence des données aberrantes éventuelles en fin d'analyse).

Exemple 1.6.1.3.2 : Soit la série des données déjà ordonnée suivante :
8 12 20 27 30 32 35 36 40 40 40 40 41 42 45 47 50 52 61 89 101.
(n=21 observations). Déterminer s'il y a des données aberrantes dans cette série à l'aide d'un graphique en boîte (box-plot).
Réponse : Les différentes mesures de cette variable sont obtenues à l'aide du logiciel Minitab:

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C2	21	0	42.29	4.72	21.63	8.00	31.00	40.00	48.50	101.00

Ce qui signifie que $Q_1 = 31; Q_3 = 48,5$ et donc $L_{inf} = Q_1 - 1,5 * (Q_3 - Q_1) = 4,75$ et $L_{sup} = Q_3 + 1,5 * (Q_3 - Q_1) = 74,75$. Donc, il y a 2 données aberrantes dans cette série ce sont 89 et 101(qui sont signalées par *), ce qui est illustré dans le diagramme en boîte ci-dessous.



Remarque : Une donnée abérrante peut avoir un effet catastophique sur la moyenne, sur l'écart type et même sur l'allure générale de la distribution des données.

1.6.2. Les autres de position.

Quelques fois, on doit découper une série d'observations en cinq, en dix ou en cents groupes contenant chacun le même pourcentage d'observations. Dans le cas de cinq groupes, on parle alors des quintiles V_1, V_2, V_3 *et* V_4. Entre deux quintiles consécutifs, il y a 20% d'observations. Dans le cas de dix groupes, on parle des déciles $D_1, D_2, \ldots, D_9$ et entre deux déciles consécutifs, il y a 10% d'observations. Dans le cas de cent groupes, on parle des centiles $C_1, C_2, \ldots, C_{99}$ et entre deux centiles consécutifs, il y a 1% des observations. Le calcul de ces différentes mesures de position est identique à ce qu'on a fait pour déterminer les quartiles, il n'y a que le pourcentage de la mesure à adapter à chaque fois. On va donner un exemple dans le cas où les données sont groupées par classes.

Exemple 1.6.2.1 : En reprenant les données de l'exemple 1.6.1.3.1, déterminer le deuxième quintile, le septième décile et le quatre vingt quinzième centile de la variable X, les recettes quotidiennes d'un petit dépanneur et interprétez chacune de ces mesures.

X=les recettes	Fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
[10 ; 100[	5	0,125	0,125
[100 ;190[	3	0,075	0,200
[190 ;280[	11	0,275	0,475
[280 ;370[	6	0,150	0,625
[370 ;460[	11	0,275	0,900
[460 ;550[	3	0,075	0,975
[550 ;640]	1	0,025	1,000
Total	n=40	1,000	

Réponse :

(a) Les fréquences cumulées dépassent pour la première fois 40% au niveau de la classe [190 ; 280[ainsi le deuxième quintile est égal à $V_2 = 190 + \frac{(0,40-0,20)}{0,275} * 90 = 255,45\$$. Ceci signifie que dans cet échantillon de données, 40% des journées, les recettes quotidiennes de ce petit magasin ont été de 255,45 $ ou moins.

(b) Les fréquences relatives cumulées dépassent pour la première fois 70% au niveau de la classe [370 ; 460[, ainsi le septième décile est égal à $D_7 = 370 + \frac{(0,70-0,625)}{0,275} * 90 = 394,55\$$. Ce qui signifie que dans cet échantillon de données, 70% des journées, les recettes quotidiennes de ce petit magasin ont été de 394,55$ ou moins.

(c) Les fréquences relatives cumulées dépassent pour la première fois 95% au niveau de la classe [460 ; 550[, ainsi le quatre vingt quizième centile est égal à $C_{95} = 460 + \frac{(0,95-0,90)}{0,075} * 90 = 520\$$. Ce qui signifie que dans cet échantillon de données, 95% des journées, les recettes quotidiennes de ce petit magasin ont été de 520$ ou moins.

1.7 : Les mesures de dispersion.

Rappelons qu'on travaille sur des données issues d'un échantillon et que le choix de cet échantillon est fait au hasard mais sensé refléter ce qui se passe dans la population. Ce qui fait que le comportement d'une variable diffère d'un échantillon à l'autre mais on espère qu'il correspond au profil de cette variable dans la population. Ce qui fait que lorsqu'on manipule une variable mesurable et qu'on se base seulement sur ses mesures de tendance centrale, on perd de vue la variabilité des données autour de ces mesures centrales. D'où l'utilité des mesures de dispersion qui, jumulées avec les mesures de tendance centrale, vont nous donner une idée plus exacte sur l'ensemble de ce qu'on a observé dans une série échantillonnale. Dans ce paragraphe, on va décrire quelques unes de ces mesures de dispersion.

1.7.1 : L'étendue.

C'est la mesure de dispersion la plus simple à calculer. Lorsqu'on a une variable quantitative X, mesurée sur un échantillon de taille n. Alors l'étendue est égale à $E = la\ plus\ grande\ donnée - la\ plus\ petite\ donnée = X_{max} - X_{min}$.
Puisque l'étendue est basée seulement sur les deux observations extrêmes, alors elle est très peu utilisée dans les applications.

1.7.2 : La variance.

La variance d'une variable mesurée sur un échantillon est égale à la moyenne des carrés des écarts qui séparent chaque observation de la moyenne échantillonnale, son calcul diffère selon la nature des données.

1.7.2.1 : Les données en vrac.

Soit X une variable quantitative mesurée sur un échantillon de taille n, et dont les valeurs sont : $x_1, x_2, \dots, x_n$ alors la variance de l'échantillon est

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

La sommation ci-dessus est divisée par (n-1) pour que cette variance échantillonnale soit une bonne estimation de la variance de toute la population. Ce qu'on verra plus en détails dans le chapitre VI. La variance se prête mal à l'interpétation car vue son calcul, son unité est égale au carré de l'unité de la variable X. Si par exemple X est égal au nombre d'enfants par ménage alors l'unité de la variance serait $(nombre\ d'enfants)^2$ qui n'a aucune signification.
La variance est surtout utile lorsqu'on a une variable mesurée dans plusieurs groupes (analyse de la variance) ou dans le cas où on veut comparer plusieures variables mesurées sur le même échantillon ou comme étape de calcul pour calculer d'autres mesures.

Exemple 1.7.2.1.1 : Soit X une variable quantitative mesurée sur un échantillon de taille n=6 et les valeurs suivantes ont été obtenues : -2 5 10 7 8 8
Alors $\bar{x} = 6$ et la variance de cet échantillon sera égale à

$$s_X^2 = \frac{(-2-6)^2 + (5-6)^2 + \cdots . + (8-6)^2}{6-1} = 18.$$

1.7.2.2 Les données groupées par valeurs.

Soit X une variable quantitative mesurée sur un échantillon de taille n, et dont les k valeurs sont : $x_1, x_2, \dots, x_k$ avec des fréquences absolues respectivement égales à $f_1, f_2, \dots, f_k$. Alors la variance de X dans cet échantillon est égale à

$$s_X^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n-1}.$$

Exemple 1.7.2.2.1 : En reprenant le tableau de l'exemple 1.5.2.2.1, déterminer la variance de la variable X=le nombre d'accidents par semaine.

Tableau des fréquences du nombre d'accidents par semaine	
X	Fréquences absolues
0	4
1	2
2	10
3	7
4	10
5	4
6	3
Total	n=40

Réponse : On avait trouvé que la moyenne de cette variable est $\bar{x} = 3{,}025$ donc sa variance sera égale à :

$$s_X^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n-1} = \frac{(0-3{,}025)^2 * 4 + \cdots + (6-3{,}025)^2 * 3}{39}$$
$$= 2{,}74\ (\frac{accidents}{semaine})^2\ !!!!$$

1.7.2.3 : Les données groupées par classes.

Soit maintenant X, une variable quantitative mesurée sur un échantillon de taille n, et dont les observations sont groupées en k classes avec des fréquences absolues respectivement égales à $f_1, f_2, \ldots, f_k$ et dont les milieux des classes sont respectivement égaux à $m_1, m_2, \ldots, m_k$. Alors la variance échantillonnale de cette variable est :

$$s_X^2 = \frac{\sum_{i=1}^{k}(m_i - \bar{x})^2 f_i}{n-1}$$

Exemple 1.7.2.3.1 : En reprenant les données de l'exemple 1.5.2.3.1, déterminer la variance de la variable X, les recettes quotidiennes d'un petit dépanneur.

Réponse : On avait trouvé que la moyenne de la variable est $\bar{x}$ =298$

m_i	X=les recettes	Fréquences absolues
55	[10 ; 100[	5
145	[100 ;190[	3
235	[190 ;280[	11
325	[280 ;370[	6
415	[370 ;460[	11
505	[460 ;550[	3
595	[550 ;640]	1
	Total	n=40

Alors la variance de cet échantillon est égale à :

$$s_X^2 = \frac{\sum_{i=1}^{k}(m_i - \bar{x})^2 f_i}{n-1} = \frac{(55-298)^2 * 5 + \cdots + (595-298)^2 * 1}{39}$$
$$= 20021{,}54\ (\$)^2 !!!!!$$

1.7.3 : L'écart type.

L'écart type d'une variable quantitative mesurée sur un échantillon est égal à la racine carrée de sa variance. Son unité de mesure étant la même que celle de la variable, l'écart type se prête alors aisément à l'interprétation et est considéré comme la mesure de dispersion par excellence. La variance n'est donc qu'une étape de calcul pour déterminer l'écart type, quand on faisait les calculs à la main. Maintenant que tout est programmé, aucune calculatrice et aucun logiciel ne parle de variance comme telle.

Exemple 1.7.3.1 : L'écart type échantillonnal pour les 3 précédents exemples où on a calculé les variances échantillonnales est respectivement égal à :

$s_X = \sqrt{18} = 4{,}24$. Pour les données de l'exemple 1.7.2.1.1 où les données sont en vrac.

$s_X = \sqrt{2{,}74} = 1{,}655$. Pour les données de l'exemple 1.7.2.2.1 où les données sont en groupées par valeurs.

$s_X = \sqrt{20021{,}54} = 141{,}497$. Pour les données de l'exemple 1.7.2.3.1 où les données sont groupées par classes.

Interprétation de l'écart type échantillonnal.

L'écart type mesure la dispersion entre toutes les valeurs observées. Des valeurs proches donneront un plus petit écart type, alors que des données très séparées donneront un plus grand écart type.

Lorsque la distribution des données (histogramme ou polygone des fréquences ou autre) a une forme en cloche et que la taille de l'échantillon est supérieure à 100, on doit s'attendre à avoir 68% des données observées comprises entre la moyenne plus ou moins un écart type et 95% des données observées soient comprises entre la moyenne plus ou moins deux écarts types. Si on se trouve dans les mêmes conditions on peut estimer l'écart type par la formule suivante :

$$s_X \approx \frac{\text{Étendue de } X}{4}$$

1.7.3.1 : Propriétés de l'écart type échantiollonnal.

Soit X une variable quantitative dont l'écart type échantillonnal est s_X et soit Y une autre variable quantitative telle que $Y = a + b * X$ où a et b sont des constantes réelles. Alors l'écart type échantillonnal de Y sera égal à

$s_Y = |b| * s_X$

Exemple 1.7.3.1.1 : Reprenons le contexte de l'exemple 1.5.2.4.1, où X est le nombre d'heures qu'un étudiant travaille à temps partiel par semaine. Supposons

qu'à partir d'un échantillon d'étudiants, on ait pu trouvé que l'écart type du nombre d'heures travaillées par ces étudiants est égal à $s_X = 3,2$ heures/semaine. Si le salaire horaire est de 10$ et que les patrons de ces étudiants leur offrent 30$ par semaine pour leurs déplacements, quel est l'écart type du gain net hebdomadaire de ces étudiants ? Posons Y, le gain net hebdomadaire de ces étudiants alors $Y = 30 + 10 * X$, donc l'écart type du gain net de cet échantillon d'étudiants sera égal à $s_Y = 10 * s_X = 32$ $/semaine.

1.7.4 : Le coefficient de variation.

On avait dit que l'unité de l'écart type d'une variable est la même que celles des données et qu'alors il s'interprète mieux que la variance. Mais si on veut comparer la dispersion de deux variables ou plus ayant des unités différentes mesurées sur le même échantillon ou sur des échantillons différents, il nous faut une mesure de dispersion sans unité. Cette mesure est le coefficient de variation. Pour un échantillon de données dont la moyenne est non négative, on définit le coefficient de variation d'une variable X par :

$$CV_X = \frac{s_X}{\bar{x}} 100\%$$

Si on a un seul échantillon de données, alors si le coefficient de variation de X est inférieur à 15%, on dit que la variable est homogène, sinon elle est dite hétérogène.

Si on a deux échantillons (sur une ou deux variables) ou plus, alors celui (ou celle) qui a le plus petit coefficient de variation est le (ou la) plus homogène.

Exemple 1.7.4.1 : On a pris un échantillon de taille n=50 d'hommes d'âge adultes, on a mesuré leur poids et leur taille. Les résultats sont résumés dans le tableau suivant :

Variable	Moyenne	Écart type
X=taille	$\bar{x}$=173,59 cm	$s_X = 7,86\,cm$
Y=poids	$\bar{y} = 78,42\,kg$	s_Y=11,98 kg

Pour comparer l'homogénéité de ces deux variables, on utilise leur coefficient de variation.

$$CV_X = \frac{7,86}{173,59} 100\% = 4,53\%$$

$$CV_Y = \frac{11,98}{78,42} 100\% = 15,28\%$$

Donc la taille des hommes adultes est plus homogène que leur poids. Ce qui correspond à l'intuition. Par exemple il est très rare de voir deux hommes

adultes dont l'un serait deux fois plus grand que l'autre, alors qu'il est fréquent de voir un homme adulte dont le poids est le double d'un autre.

Exemple 1.7.4.2 : Pour comparer les distributions des blessures graves dans le basketball et dans le soccer, on a sélectionné au hasard 25 cégeps où ces sports se pratiquent en sport-étude. On a obtenu chez les étudiants masculins, les données relatives aux nombres de blessures graves par année dans ces deux sports :

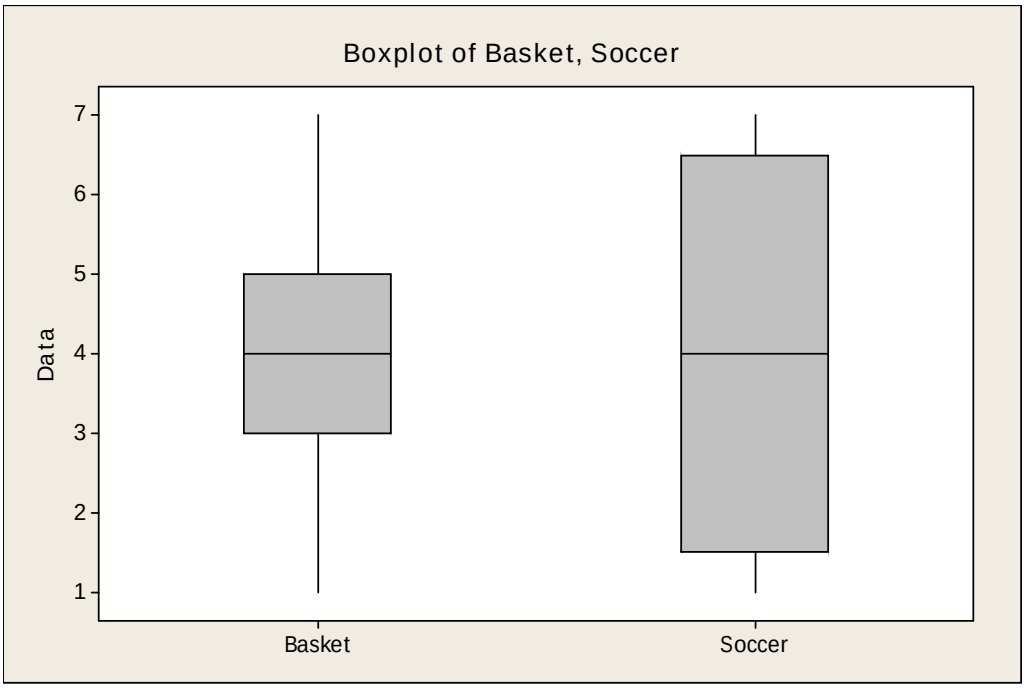
Basketball
1 2 4 4 7 3 3 2 4 5
2 4 3 5 3 4 4 3 6 5
5 6 4 6 5

Soccer
1 7 7 6 1 2 6 1 7 2
1 3 2 7 5 6 1 7 4 1
5 7 6 3 2

Pour comparer ces deux échantillons, calculons d'abord leurs mesures statistiques de base.

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Basket	25	0	4.000	0.294	1.472	1.000	3.000	4.000	5.000	7.000
Soccer	25	0	4.000	0.490	2.449	1.000	1.500	4.000	6.500	7.000

On voit que leur moyenne et leur médiane sont toutes égales à 4, donc si on se limitait aux mesures de tendances centrales, on aurait conclu à une similitude de ces deux distributions.
Mais en comparant leur écart type et donc leur coefficient de variation, on voit que les données sur le soccer sont plus dispersées. Ce qu'on peut aussi confirmer par des graphiques suivants :



Ayez un esprit critique.

Biostatistique L_3

Maintenant qu'on est armé d'outils pour examiner la tendance centrale, la dispersion, la distribution des données, les valeurs extrêmes ou aberrantes, on pourrait être tenté de développer une procédure mécanique et aveugle, mais penser de façon critique est d'une importance primordiale dans toute analyse de données. En plus de l'utilisation des outils présentés dans ce chapitre, il est important de ne pas négliger tout autre facteur qui s'y rapporte et qui pourrait être crucial pour les conclusions de l'étude. On pourrait penser par exemple à la représentativité des données, à la source des données qui pourrait affecter leur qualité. En résumé, en plus des outils présentés dans ce chapitre, on devrait aussi penser.