www.fundp.ac.be/biostats **Module 10**

10Variabilité et statistiques

10.1 Population et mesures expérimentales

On peut définir la statistique comme la science du dépouillement des données numériques fournies par l'observation d'un phénomène naturel.

> *Depuis plusieurs années, l'administration des eaux et forêts importe de Bohème (Tchequie) deux espèces de poissons :* Coregonus peled *(Gmelin, 1788) et* Coregonus lavaretus *(Linné, 1758), en français le peled et le lavaret, pour réempoissonner le lac artificiel de Robertville. La question se pose de déterminer laquelle des deux espèces s'adapte le mieux à cet environnement.*

Cette étude va donc être réalisée dans une certaine **population** (les poissons du lac) dans laquelle nous allons observer le phénomène par certaines **mesures expérimentales.**

> *Nous devons donc définir cette mesure : on peut par exemple choisir de mesurer la taille des poissons d'un an, exprimée en cm. C'est ce que nous appellerons la **variable**. Le choix de la variable sera fort important pour la suite des opérations. Notons que dans cet exemple, nous aurions pu décider de mesurer le poids[1] du poisson (en g)*

Cette mesure expérimentale ne peut généralement pas être effectuée sur tous les individus de la population.

> *Dans ce cas -ci, il est illusoire de prétendre capturer tous les poissons d'un an vivant dans le lac. Un recensement total ne s'obtiendrait que par un empoisonnement total ou la vidange du lac : il faut reconnaître que cette technique, à laquelle on recourt parfois, présente un certain nombre d'inconvénients.*

On devra donc se limiter à quelques poissons capturés : la fraction de la population pour laquelle la mesure expérimentale aura été effectivement réalisée s'appelle **l'échantillon,** et la technique utilisée pour le récolter, **l'échantillonnage**.

[1] Dans cet ouvrage, nous utilisons systématiquement poids pour masse corporelle, ce qui correspond au langage usuel, bien que scientifiquement, le poids se mesure en Newton et la masse en g ou kg.

Phénomène naturel	croissance du poisson
Variable Mesure expérimentale	taille à un an (en cm) réglette graduée en mm
Population	Coregonus peled individus d'un an, Lac de Robertville, année 2002
Modèle :	La croissance du poisson dépend de l'âge
Echantillonnage Technique Observation complémentaire	aléatoire pêche au filet identification de l'âge des poissons
Echantillon	n peleds d'un an
Série statistique	n valeurs exprimées en cm
Valeurs individuelles	x_{11}, x_{12}, ... x_{1i}, ... x_{1n1}

Tableau 10-1 Etapes menant à l'étude d'un phénomène naturel par la constitution d'une série statistique.

Nous disposerons ainsi d'une série de valeurs numériques (en cm) :

13,2 ; 12,4 ; 11,9 ; 14,3 ; 10,4 ...

que nous noterons, en toute généralité :

x_1, x_2, x_3 ...

ce sont les **valeurs individuelles**. L'ensemble des ces valeurs constitue **la série statistique.**

Ce que notre pêcheur va constater en tout premier lieu est que les peleds d'un an n'ont pas tous la même taille. Il est confronté à un *problème de variabilité*.

10.2 En quoi consiste la variabilité?

Différentes mesures donnent des résultats différents pour trois grandes raisons. Illustrons ceci en imaginant un ou plusieurs tireurs effectuant des tirs à la carabine sur une cible fixe :

10.2.1 Imprécision

Imaginez qu'un individu tire deux coups consécutifs avec la même carabine.

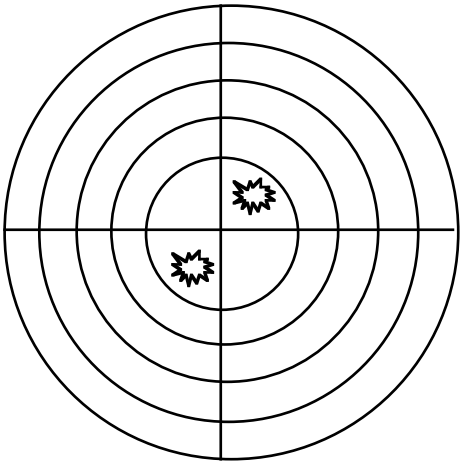

Figure 10 -1 Impacts de deux tirs précis réalisés par un seul tireur utilisant une seule carabine bien réglée.

La distance entre deux impacts montre que dans des circonstances semblant absolument identiques, le résultat obtenu n'est pas identique. Nous observons ici une forme de la variabilité "irréductible", qui doit être rendue la plus petite possible, mais qu'il est vain de penser pouvoir supprimer. Cette variabilité fait partie de ce que l'on appellera la variabilité *résiduelle*.

Cette variabilité est celle que l'on observera en mesurant, ou en pesant deux fois de suite un même poisson : il ne sera pas étiré sur la latte exactement de la même façon, pas épongé exactement de la même façon avant la pesée…

Il faut noter que cette variabilité n'est en rien une erreur, bien que dans certains textes ou logiciels on la désigne abusivement comme erreur. Il s'agit d'une imprécision qui est la résultante de toute une série de phénomènes incontrôlés (voir plus bas : école déterministe et stochastique).

Dans le second tir, nous pouvons imaginer que le viseur de la carabine est déréglé, car la distance entre les deux impacts (imprécision) apparaît très petite par rapport à la distance entre les impacts et le centre de la cible :

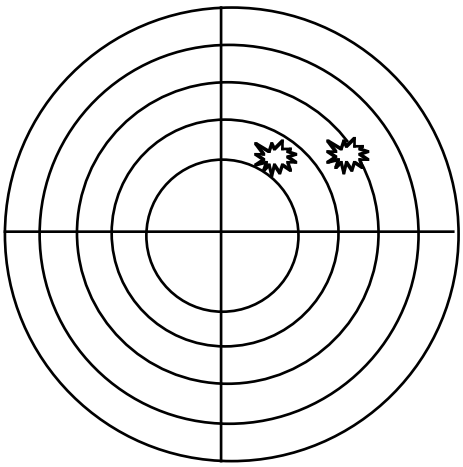

Figure 10 -2 Impacts de deux tirs précis réalisés par un seul tireur, utilisant une seule carabine mal réglée.

Cette différence systématique entre la mesure effective (ici l'impact) et la valeur réelle (ici le centre de la cible) représente l'inexactitude.

On parlera aussi d'erreur, cette fois dans le vrai sens du terme : je peux imaginer que la balance est mal tarée, ou que le poisson est mal égoutté ; il peut arriver de renverser inopinément une partie d'un liquide avant d'en mesurer le volume, de se tromper dans le réglage d'une pipette : en cas d'erreur (inexactitude) il y a une différence systématique entre la mesure et la réalité.

Remarquons que si l'imprécision est grande, elle pourra masquer l'inexactitude : lors de ce troisième tir, nous avons déréglé le viseur de la même façon que pour le tir précédent :

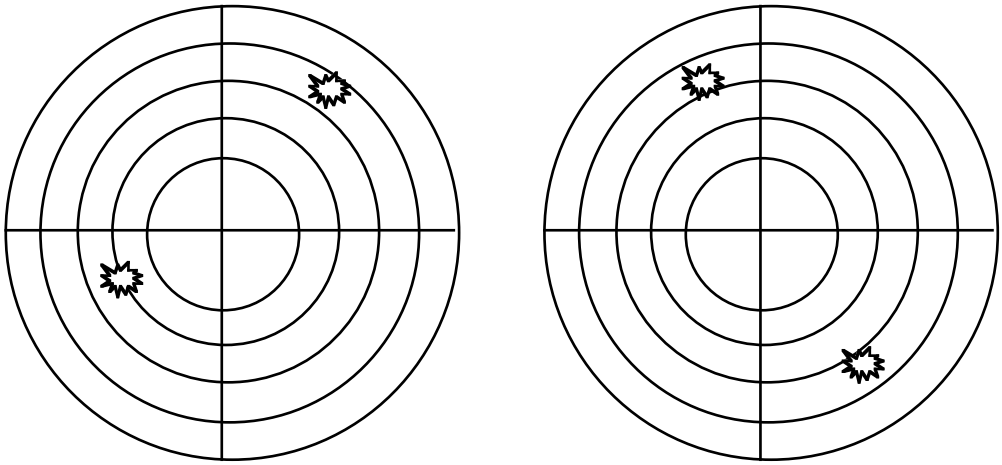

Figure 10 -3 Impacts de deux tirs imprécis réalisés par un seul tireur utilisant une seule carabine, bien réglée (cible de gauche) et mal réglée (cible de droite).

le résultat obtenu ne permet cependant pas de le discerner car la distance entre les deux impacts (imprécision) est aussi grande que la distance entre le centre de la cible et les impacts (inexactitude).

Une différence entre l'observation et la réalité peut également résulter d'un biais dans l'échantillonnage. Par exemple, si par facilité tous les poissons sont pêchés à partir du rivage, ils sont peut -être systématiquement plus petits que les poissons de la même population qui vivent au large : la taille sera dès lors sous -estimée.

Dans tout ce qui suit, nous considérerons que

- la mesure, bien qu'imprécise, est toujours exacte ;

- l'échantillon est représentatif de la population, c -à -d que tout individu de la population a la même chance de se trouver dans l'échantillon.

10.2.2 Différences individuelles

Imaginons maintenant que deux individus tirent sur la cible, avec la même carabine. Nous voyons apparaître une nouvelle forme de variabilité. Le premier tireur (impacts noirs) n'a pas la même déviation par rapport au centre de la cible que le second tireur (impacts blancs).

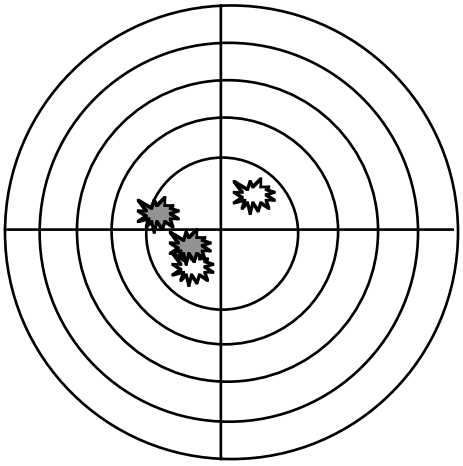

Figure 10 -4 Impacts de deux tirs précis réalisés par deux tireurs utilisant une même carabine bien réglée.

Cette forme de variabilité est celle que l'on observera en mesurant, ou en pesant, des peleds d'un an vivant exactement dans les mêmes conditions. Elle caractérise la population, et se reflète dans l'échantillon. Aussi précise que soit la mesure, elle affectera la série statistique. Comment définir le « poids réel » d'un individu lorsque celui -ci change continuellement de poids ?

Si la mesure n'est effectuée qu'une seule fois par individu, la variabilité individuelle sera indissociable de l'imprécision, car la seule mesure de distance disponible est à la fois la distance entre deux répétitions de la mesure et entre deux individus différents.

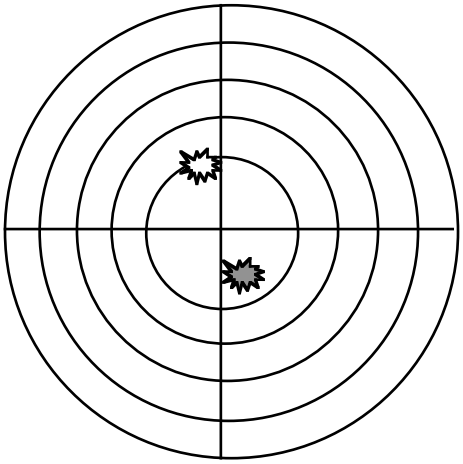

Figure 10 -5 Impacts d'un seul tir réalisé par deux tireurs utilisant une même carabine bien réglée.

Dans ce cas, ce que l'on appelle la variabilité résiduelle sera la somme de la variabilité individuelle (distance entre les valeurs réelles de différents individus) et de l'imprécision de la mesure (distance entre les valeurs mesurées pour un même individu)

Il ne faut pas confondre variabilité individuelle et individu « hors norme » (outsider). Il se peut que pour des raisons parfois inexpliquées, un individu présente une valeur exceptionnelle, peut -être à cause d'une malformation. Dans ce cas, il ne fait en réalité pas partie de la population que l'on souhaite caractériser. Bien que la prudence s'impose, de telles observations sont généralement retirées de la série statistique.

Dans ce qui suit, nous considérerons que les individus considérés sont représentatifs de la population.

10.2.3 Différences factorielles

Imaginons maintenant que deux individus tirent sur la cible avec la même carabine, puis chacun recommence avec une autre carabine. Nous voyons apparaître une nouvelle forme de variabilité.
La première carabine (impacts circulaires) n'a pas la même déviation par rapport au centre de la cible que la seconde carabine (impacts carrés). Cette différence est due à un facteur identifiable, qui est ici l'emploi de deux carabines différentes.

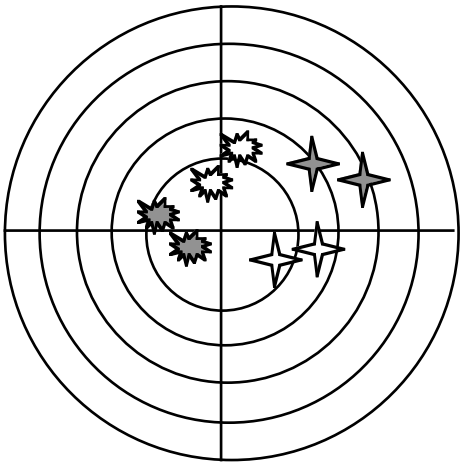

Figure 10 -6 Impacts de deux tirs réalisés par deux tireurs (blanc, noir), utilisant successivement une même carabine bien réglée (ronds), puis une même carabine mal réglée (étoile).

Dans le cas de nos poissons, il s'agirait de différence de mesures entre peleds et lavarets (facteur espèce). On pourrait imaginer qu'il s'agit de différences entre peleds d'un an et de deux ans (facteur âge), entre peleds mâles et femelles (facteur sexe)... C'est ce que l'on appellera la variabilité factorielle.

Cette variabilité se distingue fondamentalement de la variabilité résiduelle : elle représente de l'information qui permet de répondre à des questions correspondant aux différents niveaux du facteur : les carabines sont -elles réglées de la même façon (niveaux 1 et 2 du facteur carabine)? Les peleds atteignent -ils la même taille à un an que les lavarets (niveaux 1 et 2 du facteur espèce) ? Quelle est la différence de taille entre un et deux ans (niveaux 1 et 2 du facteur âge)? Quelle est la différence de taille entre mâles et femelles (niveaux 1 et 2 du facteur sexe) ?

> *La variabilité résiduelle peut être considérée comme du bruit parasite : si c'était possible, je préférerais que le poids affiché sur la balance soit exactement la valeur réelle du poids de l'individu, et que deux alevins placés dans les mêmes conditions pendant 365 jours aient exactement le même poids: je pourrais ainsi répondre directement à la question !*

Notons que si je ne dispose pas, au minimum, de la mesure de deux individus dans chacune des conditions, la variabilité factorielle sera indissociable de la variabilité résiduelle.

En effet, la seule mesure de distance disponible est à la fois la distance entre deux individus de la mesure et des niveaux du facteur différents.

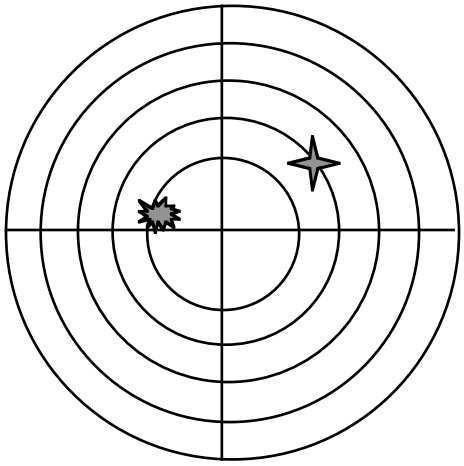

Figure 10 -7 Impacts de deux tirs réalisés par un tireur utilisant successivement deux carabines (rond et carré).

Pour faire apparaître la variabilité factorielle, le facteur doit être identifié et les observations individuelles regroupées par niveaux de ce facteur. Par exemple, si je mesure la croissance des poissons sans en identifier le sexe, la variabilité entre mâles et femelles viendra s'ajouter à l'imprécision et à la variabilité individuelle et l'ensemble formera la variabilité résiduelle.

L'imprécision peut être détectée si j'effectue la mesure au moins deux fois, dans les mêmes conditions, sur le même individu (ces mesures sont appelées duplicats).

La variabilité individuelle peut être détectée si j'effectue la mesure, dans les mêmes conditions, sur au moins deux individus (ces mesures sont appelées réplicats).

La variabilité factorielle peut être détectée si j'effectue la mesure, dans les mêmes conditions, sur au moins deux individus et pour au moins deux niveaux de ce facteur.

Phénomène naturel		croissance du poisson	
variable **mesure expérimentale**		taille à un an (cm) réglette graduée en mm	
Modèle :		La croissance du poisson dépend de l'espèce et de l'âge	
	population 1 Coregonus peled individus d'un an, Lac de Robertville, année 2002		population 2 Coregonus lavaret individus d'un an, Lac de Robertville, année 2002
échantillonnage sélection des observations observations complémentaires		aléatoire pêche au filet identification des espèces identification de l'âge des poissons	
	Echantillon 1 n1 peleds d'un an		Echantillon 2 n2 lavarets d'un an
série statistique		exprimées en cm	
	valeurs individuelles x_{11}, ... x_{1i}, ... x_{1n1}		valeurs individuelles x_{21}, ... x_{2i}, ... x_{2n2}

Tableau 10-2 Etapes menant à l'étude d'un facteur à deux niveaux sur un phénomène naturel par la constitution de deux séries statistiques.

La variabilité résiduelle est la variabilité qui ne peut pas être dissociée de l'imprécision, par manque de duplicats, de réplicats, ou d'identification du facteur.

La variabilité totale observée est la somme de la variabilité factorielle et de la variabilité résiduelle.

10.3 Comment maîtriser la variabilité?

10.3.1 Pourquoi la mesure est -elle variable?

Deux approches complémentaires permettent de donner des éléments de réponses. Selon **l'école déterministe**, un ensemble complexe de mécanismes précis génère une impression de réponse aléatoire. La météorologie en est un bon exemple : mieux les mécanismes déterminant les variations du temps sont connus et font l'objet de mesures précises, plus les prédictions sont précises.

Dans notre exemple, nous pouvons considérer que la taille du poisson d'un an est la résultante de l'influence d'une série de facteurs génétiques, de la disposition de nourriture, de l'énergie qu'il dépense à se nourrir, du rendement de son métabolisme, etc...

Suivant l'école déterministe, si nous avions les moyens d'identifier et de quantifier toutes ces sources, nous pourrions déterminer avec exactitude la taille d'un poisson donné.

Selon l'**école stochastique**, un mécanisme comprenant au moins une étape aléatoire génère une réponse aléatoire.

Certains mécanismes biologiques, telle la sélection des allèles à la méiose ou la création du répertoire de spécificité des lymphocytes T & B (défense immunologique) relèvent d'un hasard comparable à celui d'un jeu de dés!

En suivant l'école stochastique, il n'est pas possible de prédire la taille du poisson d'un an autrement que par une estimation du type : *il est fort probable qu'un poisson d'un an ait une taille comprise entre certaines limites*.

10.3.2 En quoi la variabilité pose -t -elle un problème?

peled	lavaret
11,2	15,3
12,2	15,0
10,0	17,2
11,1	15,5
11,0	14,5
11,4	14,5
14,3	14,0
11,4	14,6
11,4	15,2

Tableau 10-3 Capture de poissons qui montre sans ambiguïté une meilleure croissance des lavarets,

Si mon travail est de déterminer quelle espèce (peleds ou lavarets) a la meilleure croissance dans les eaux du lac de Robertville, la variabilité risque de m'empêcher de prendre une décision claire : imaginons que la pêche produise les résultats de la table 1.3. : le bon sens suffirait à nous convaincre que les lavarets ont une meilleure croissance que les peleds.

Une étude statistique plus détaillée pourrait prouver que l'échantillon est assez large pour qu'un tel résultat ne puisse pratiquement jamais être observé si les deux populations avaient la même croissance, mais elle n'est pas indispensable.

Par contre, devant les résultats de la table 1.4, il serait impossible de formuler intuitivement la même affirmation. Pourtant, une étude statistique plus détaillée pourrait prouver que les lavarets ont ici aussi une meilleure croissance que les peleds !

peled	lavaret
13,6	11,9
12,2	12,9
10,2	11,9
12,4	13,1
11,2	12,2
12,6	12,8
12,6	13,1
11,2	13,5
10,9	12,9

Tableau 10-4 Capture de poissons qui ne montre pas de façon évidente une meilleure croissance des lavarets.

10.3.3 En quoi les statistiques apportent -elles une solution à ce problème?

Il faut avouer qu'une affirmation du type « il est fort probable qu'un poisson d'un an ait une taille comprise entre certaines limites » s'apparente aux "réponses de Normand".

Penseriez -vous acheter une voiture d'occasion à un vendeur qui vous dirait « je pense que cette voiture pourra encore rouler un certain nombre de kilomètres » ?

Le but des statistiques est de permettre de quantifier les expressions "*il est fort probable"* et "*certaines limites".* On arrive alors par exemple à dire : 9*5% des peleds d'un an ont une taille comprise entre 10 et 14 cm*.

Ou encore : il y a 95% de chances que ce véhicule roule encore entre 40.000 et 60.000 km, ce qui est nettement plus crédible.

10.3.4 Peut -on supprimer la variabilité?

Il est légitime de se demander s'il n'est pas préférable de limiter la variabilité plutôt que de développer des techniques qui permettent de la contourner. Dans cette réflexion, il est indispensable de distinguer variabilité individuelle, imprécision et inexactitude.

L'imprécision peut - et doit - être limitée par tous les moyens techniques appropriés : choix de la variable (exemple : la taille est moins variable que le

poids), le protocole de mesure (exemple : endormir le poisson avant de le mesurer), la graduation de la latte (la balance est peut-être plus précise ?) …

la variabilité individuelle ne peut pas être supprimée. On peut cependant standardiser l'expérience, en ne mélangeant pas les âges, les sexes, les races… dans la même expérience. Il s'agit ici de limiter la définition de la population.

> *Plutôt que d'étudier « l'évolution de la consommation de tabac », étudions par exemple « l'évolution de la consommation de tabac, chez les garçons belges de 17 à 25 ans ».*

Un des enjeux du clonage (par exemple de cellules, d'embryons…) est la possibilité de réaliser des réplicats les moins variables possibles, afin de permettre de mettre en évidence la variabilité factorielle la plus petite possible (et donc des différences plus subtiles, suite à la variation d'un facteur donné).

La **variabilité factorielle** représente la variabilité induite par l'appartenance des observations à des populations différentes (mâles/ femelles, un an/deux ans…). Elle est donc, par définition, irréductible.

> *D'ailleurs cette variabilité est souvent induite par l'expérimentateur, pour éprouver une hypothèse de travail. Par exemple il voudrait montrer l'influence de la température sur la croissance du poisson. Pour cela il induit une différence de croissance en élevant des poissons dans des bassins chauffés à 20 et 25°C.*

La variabilité représente dès lors une information et non plus un bruit parasite. Le but est alors de rendre ce signal

le plus intense possible en maximisant l'effet expérimental
le plus perceptible possible en réduisant la variabilité résiduelle.

> *Si vous souhaitez entendre les nouvelles à la radio (ce qui représente l'information : variabilité factorielle) alors que le bruit extérieur vous en empêche (bruit qui représente la variabilité résiduelle) vous pouvez soit augmenter le volume de la radio (ce qui représente : maximiser l'effet expérimental : comparons 10 et 25°C) soit fermer la fenêtre (ce qui représente diminuer le bruit : expérimentons seulement les femelles d'un an).*

10.4 Paradoxe fondamental

Nous pouvons donc découvrir ici un paradoxe fondamental des statistiques :

Il peut être possible de prouver que la température a un effet sur la croissance : cela sera le cas si la variabilité factorielle est plus grande que la variabilité résiduelle (l'information est plus grande que le bruit).

Dès qu'il y a une variabilité résiduelle, il sera toujours impossible de *prouver* que la température n'a pas un effet sur la croissance. En effet, on pourra toujours suspecter que cet effet existe, mais qu'il est masqué par la variabilité résiduelle (le bruit couvre l'information).

La possibilité d'arriver à une conclusion par une technique statistique est donc toujours fonction du rapport entre la variabilité de la mesure (résiduelle) et les différences entre les populations considérées (factorielle).

Chacune de ces variabilités est composée d'une partie sur laquelle l'expérimentateur peut agir et une sur laquelle il ne le peut pas.

10.5 Mesures de la variabilité

http://www.fundp.ac.be/biostats/biostat/modules/module10/index.html - module_10

10.5.1 Ecarts à la moyenne arithmétique

Reprenons l'exemple du tir à la carabine, cette fois en retirant, après le tir, la mire de la cible et estimons par Mx,My le point que le tireur devait viser, sous réserve que les écarts au centre de la cible soient purement dus au hasard (écarts résiduels).

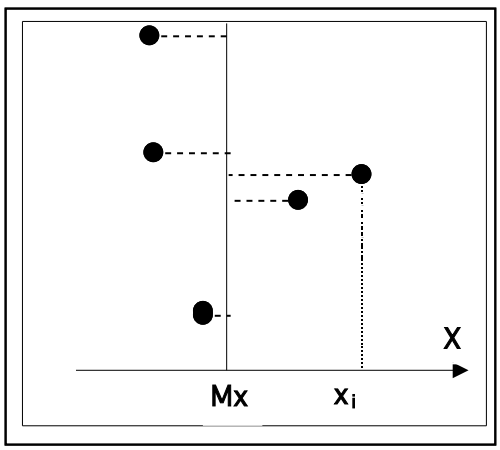

Figure 10 -8 Ecarts à la moyenne Mx, considérés sur la variable X.

Prenons en considération, dans un premier temps, seulement les écarts horizontaux (suivant un axe X). Notons x_i la coordonnée d'un impact sur X et Mx la coordonnée estimée sur X de l'axe vertical de la mire.

Chaque écart s'écrit donc
$(x_i - Mx)$

et puisque les écarts sont aléatoires, les écarts vers la gauche compensent les écarts vers la droite.

$$\sum_{i=1}^{n}(x_i - Mx) = 0$$

Équation 10-1

Ce qui s'écrit aussi :
$$\sum_{i=1}^{n} x_i - nMx = 0$$

Mx est la moyenne arithmétique des observations x_i (i = 1,…,n).

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

Équation 10-2

La moyenne arithmétique garantit que la somme des écarts à la moyenne est nulle [Equation 1-1].

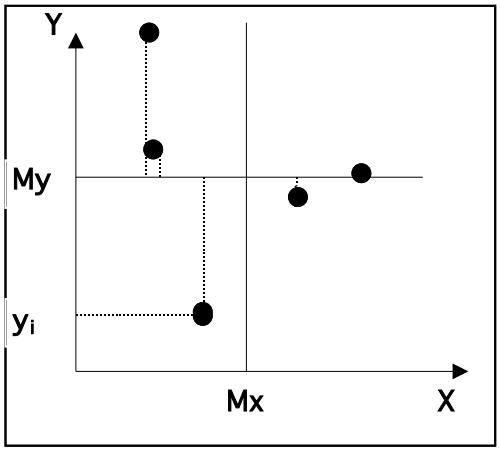

Figure 10 -9 Ecarts à la moyenne My, considérés sur la variable Y.

Prenons en considération, dans un second temps, les écarts verticaux (suivant un axe Y). Notons y_i la coordonnée d'un impact sur Y et My la coordonnée estimée sur Y de l'axe horizontal de la mire (Figure 1.9).

Nous obtenons le point Mx, My qui est la tendance centrale des impacts. La somme des écarts à la moyenne en X et en Y est nulle : les impacts sont également répartis autour de la tendance centrale.

10.5.2 Equation des sommes des carrés des écarts

Reprenons notre pêche de Coregones et supposons que nous obtenions, pour le poids de 6 alevins d'une semaine, la série statistique X (en g) :

	X g	X-Mx g	$(X-Mx)^2$ g^2
	4	-2	4
	8	+2	4
	6	0	0
	5	-1	1
	7	+1	1
	6	0	0
Mx	**6**	0	0
	Σ	**Σ**	**SCET**
	36	**0**	**10**

Tableau 10-5 Poids de 6 alevins de Coregones (g), moyenne arithmétique, écarts à la moyenne, somme, somme des écarts, carrés des écarts et somme des carrés des écarts.

La somme des valeurs est 36 g, la moyenne arithmétique 6 g. La somme des écarts (colonne 2) est nulle. Elevons ces écarts au carrés (colonne 3), puis sommons -les : la somme des carrés des écarts (SCE) est égale à 10 g^2.

De façon générale, pour la variable X :

$$\sum_{i=1}^{n} (x_i - Mx)^2 = SCEx$$

Équation 10-3

TOUTES les observations étant considérées, cette SCE est appelée la somme des carrés des écarts TOTALE (SCET) et mesure une variabilité inexpliquée (bruit).

$$Variabilité = bruit$$

A présent, considérons un modèle qui prend en compte que les alevins appartiennent à deux espèces différentes.

$$Variabilité = facteur\ espèce + bruit$$

Regroupons les peleds ensemble et les lavarets ensemble :

	Peleds	Lavarets
	4 g	6 g
	5 g	7 g
	6 g	8 g
Moyennes	**5 g**	**7 g**
Moyenne	**6g**	

Tableau 10-6 Poids individuels et moyenne arithmétique des mesures de Corégones, classés par espèce.

Les écarts que nous observons ici sont des écarts mesurés entre individus d'une même espèce, en prenant en compte le fait que les poissons appartiennent à deux espèces.

Calculons les écarts et la SCE :

	Peleds		Lavarets	
	$(x_i - M_x)$ g	$(x_i - M_x)^2$ g^2	$(x_i - M_x)$ g	$(x_i - M_x)^2$ g^2
	-1	1	-1	1
	0	0	0	0
	1	1	1	1
SCER		**2**		**2**

Tableau 10-7 Ecarts, carrés des écarts et somme des carrés des écarts des tailles de Corégones, classées par espèce.

> *Cette « prise en compte » d'un (ou plusieurs) facteur(s) s'appelle la modélisation. Le modèle identifie la variabilité qui représente une information. Les écarts qui restent sont appelés RESIDUELS car ils ne sont pas expliqués par le modèle (résidu = ce qui reste après un traitement) et la SCE est appelée somme des carrés des écarts RESIDUELS (SCER).*

Globalement, pour toutes les observations, SCER vaut 2 g^2 +2 g^2 = 4 g^2.

A présent, considérons qu'un peled vaut 5 g et un lavaret 7 g, chaque individu étant représenté par la valeur moyenne de son espèce dans l'échantillon. La moyenne globale Mx vaut donc 6 g, car il y a le même nombre d'individus de chaque espèce.

Peleds			Lavarets		
x_i	$x_i - M_x$	$(x_i - M_x)^2$	x_i	$x_i - M_x$	$(x_i - M_x)^2$
g	g	g^2	g	g	g^2
5	-1	1	7	1	1
5	-1	1	7	1	1
5	-1	1	7	1	1
SCEF		**3**			**3**

Tableau 10-8 Poids moyen par espèce, écarts à la moyenne globale, carrés des écarts et somme des carrés des écarts des tailles de Corégones, classés par espèce.

Les carrés des écarts sont ici tous égaux à 1 g^2 et la somme des carrés des écarts, cumulée pour les deux échantillons vaut 6 g^2.

Les écarts résiduels (entre individus d'une même espèce) ayant été masqués, il ne subsiste que les écarts entre espèces. Les écarts considérés ici sont donc les écarts FACTORIELS (dus au facteur espèce) et la SCE obtenue s'appelle la somme des carrés des écarts factoriels (SCEF).

Si nous considérons les trois SCE obtenues :

SCET = 10 g^2 , SCER = 4 g^2 , SCEF = 6 g^2, on découvre l'égalité 10 = 4 + 6

$$SCE_T = SCE_F + SCE_R \qquad \textbf{Équation 10-4}$$

On peut démontrer que cette équation est toujours vérifiée.

Cette équation exprime que la variabilité totale est la SOMME de la variabilité factorielle (entre les niveaux d'un facteur) et de la variabilité résiduelle (entre les réplicats pour un même niveau du facteur). Les expressions suivantes sont équivalentes :

Variabilité = information + bruit

Variabilité totale = variabilité expliquée + variabilité inexpliquée par un modèle

Variabilité totale = variabilité factorielle + variabilité résiduelle

SCEF exprime la variabilité identifiée comme interprétable par le modèle. SCER n'exprime pas d'information expliquée. Ceci ne veut pas dire qu'elle soit inexplicable : *le développement du modèle peut continuer à la*

décortiquer. Par exemple, la variabilité résiduelle contient probablement une composante liée au sexe du poisson (les mâles et les femelles ont généralement une croissance différente). Si le plan d'expérience prévoit d'identifier le sexe, et si les observations le permettent, le modèle pourra être complexifié et l'information liée au sexe pourra être aussi retirée de la SCER.

Variabilité = facteur espèce + facteur sexe + bruit

Les sommes de carrés d'écarts mesurent la variabilité entre les valeurs d'une série statistique. Elles sont exprimées dans le carré des unités. Les sommes des carrés des écarts sont additives. Les différentes composantes de la SCE permettent d'identifier différentes sources de variabilité, en fonction d'un modèle.

10.5.3 Variance

	Peleds	Lavarets
	4 g	6 g
	5 g	7 g
	6 g	8 g
		7 g
		8 g
		6 g
Mx_i	**5 g**	**7 g**
Mx	**6.33 g**	
n	**3**	**6**
$SCER_i$	**$2g^2$**	**$4\ g^2$**
SCER	**$6g^2$**	

Tableau 10-9 Valeurs individuelles des poids de deux échantillons de poisson et statistiques de ces échantillons.

Considérons une pêche de 3 peleds et de 6 lavarets (Tableau 1-9).

Nous pouvons recommencer la démarche précédente et trouver :

SCET = 14 g^2, SCEF = 8 g^2, SCER = 6 g^2

L'équation [1-4] étant toujours vérifiée.

Toutefois, une information n'est pas clairement identifiée ici : la variabilité des peleds et des lavarets est la même (les individus s'écartent tous de ± 1 par rapport à la moyenne) et pourtant, SCER = 2 g^2 dans un groupe et SCER = 4 g^2 dans l'autre groupe.

Cette différence est tout à fait logique puisque la somme des carrés des écarts est effectuée sur 3 observations dans un groupe et sur 6 observations dans l'autre groupe.

Pour exprimer que la variabilité des peleds et des lavarets est la même, il faut faire la moyenne des SCE : cette mesure s'appelle variance, ou écart quadratique moyen :

$$S_n^2 = \frac{SCE_R}{n}$$ **Équation 10-5**

On trouve dès lors

n	3	6
SCER	**2** g^2	**4** g^2
CMR ou S^2	2/3 g^2	2/3 g^2

Tableau 10-10 Statistiques du poids des peleds et lavarets : n= taille de l'échantillon, SCER = somme des carrés des écarts résiduels, CMR ou S^2 = variance résiduelle.

La variance est une mesure de la variabilité moyenne, exprimée dans le carré des unités. Les variances ne s'additionnent pas.

L'expression « variance », généralement notée S^2, peut prêter à confusion. En effet, nous justifierons plus tard une autre expression de la variance :

$$S_{n-1}^2 = \frac{SCE_R}{n-1}$$ **Équation 10-6**

Retenons que l'expression [1-5] décrit la variance d'un échantillon, tandis que l'expression [1-6] décrit la variance de la population estimée par l'échantillon. La terminologie, la notation, les symboles d'un texte, d'une calculette ou d'un tableur ne mettent pas toujours clairement en évidence si l'on se réfère à l'une ou à l'autre définition. L'expression « écart quadratique moyen » est plus logique pour désigner l'équation 1-5, mais elle n'est pas courante.

10.5.4 Ecart -type

Le fait que la variance soit exprimée dans le carré des unités limite son interprétation. Pour revenir dans les unités d'origine, on calcule la racine carrée de la variance, qui s'appelle écart -type ou déviation standard (S_n pour indiquer n observations ou Sx pour indiquer qu'il s'agit de la variable X).

$$S_n = \sqrt{\frac{SCE_R}{n}} = \sqrt{S_n^2}$$ **Équation 10-7**

Il ne faut pas s'attendre à ce que les observations soient comprises en la moyenne ± un écart -type : cet intervalle représente l'écart moyen des observations. On montrera que dans le cas de distributions symétriques, la majorité des observations se situent dans l'intervalle : moyenne ± 2 écarts – types.

n	3	6
Mx	5 g	7 g
Sx	0,82 g	0,82 g

Tableau 10-11 Statistiques du poids des peleds et lavarets : n taille de l'échantillon, Mx = moyenne, Sx = écart -type.

L'équation [1-7] décrit l'écart-type de l'échantillon et l'équation [1-8] celui de la population. Les remarques de notation énoncées pour la variance sont valables pour l'écart -type.

$$S_{n-1} = \sqrt{\frac{SCE_R}{n-1}} = \sqrt{S_{n-1}^2}$$ **Équation 10-8**

10.5.5 Propriétés algébriques de la moyenne et de la variance

Les changements de systèmes d'unités correspondent à des transformations de variables qui vont également transformer la moyenne et la variance. Par exemple, le poids peut être transformé de g en kg. Cette transformation revient à multiplier toutes les valeurs expérimentales par 0,001. On pourrait aussi devoir transformer des degrés Centigrades en degrés Kelvin, en ajoutant 273 à toutes les valeurs. Nous envisageons la transformation générale :

$X'' = a + bx$

Cette transformation permet d'ajouter à X une constante **a**, ou de multiplier X par une constante **b**, ou les deux à la fois, ce qui permet tous les changements d'unités.

*Elle est appelée transformation **linéaire** car si vous mettez sur un graphique les valeurs de x en abscisse et de x'' en ordonnée, le graphe représentera une droite (ce qui n'est pas le cas pour une transformation logarithmique ou racine carrée, par exemple).*

Les changements d'unités sont fréquents : nous allons voir comment transformer la valeur de la moyenne et de la variance de la variable x'', sans devoir retourner à la valeur des observations individuelles.

$$x_i'' = a + bx_i$$

$$Mx'' = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) =$$

$$\frac{1}{n}na + b\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$= a + bMx$$

Ceci montre que la même transformation s'applique à la moyenne des observations : si vous multipliez X par 1000, vous multipliez Mx par 1000 ; si vous ajoutez 273 à X, vous ajoutez 273 à Mx .

$$x_i'' = a + bx_i$$

$$Mx'' = a + bMx$$

$$S^2_{x''} = \frac{1}{n}\sum_{i=1}^{n}(x''_i - Mx'')^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - bMx)^2 =$$

$$b^2\frac{1}{n}\sum_{i=1}^{n}(x_i - Mx)^2$$

$$= b^2 S_x^2$$

Ce qui montre que le facteur **b** intervient au carré, mais que le facteur **a** n'intervient pas dans la transformation : si vous multipliez X par 10, vous multipliez S^2_X par 100 ; si vous ajoutez 273 à X, vous ne modifiez pas S^2_X .

1 - Si l'on ajoute ou retire une constante aux observations, on ajoute ou retire cette constante de la moyenne des observations, sans modifier leur variance.

En effet, la distribution des valeurs est simplement déplacée vers la gauche ou vers la droite, mais sa dispersion reste la même.

2 - Si l'on multiplie ou divise les observations par une constante, on multiplie ou divise la moyenne par cette constante, et la variance par le carré de cette constante.

Ceci montre que la variance et l'écart -type sont des mesures absolues de la dispersion, liées au système d'unité employé.

10.5.6 Coefficient de variation

Une mesure relative de l'écart -type, par rapport à la moyenne est donnée par le coefficient de variation :

$$C.V. = \frac{Sx}{Mx}$$ Équation 10-9

Imaginons que nous pesions des peleds à une semaine, puis à un an :

	Peleds	
	1 semaine	1 an
x_i	3 g	30 g
	3 g	30 g
	5 g	50 g
	5 g	50 g
Mx	4 g	40 g
Sx	1 g	10 g
C.V.	0,25	0,25

Tableau 10-12 Statistiques du poids des peleds et lavarets : Mx = moyenne, Sx = écart -type, C.V. = Sx/Mx = coefficient de variation.

Le coefficient de variation exprime qu'ici l'écart -type représente 25% de la moyenne.

Le coefficient de variation exprime le rapport entre l'écart -type et la moyenne. Il est dépourvu d'unité.

Dans un changement d'unité, C.V.
- n'est pas modifié par la multiplication par une constante ;
- est modifié par l'addition d'une constante.

Le coefficient de variation est souvent utilisé pour exprimer la précision d'un appareil. Si l'on dit qu'une balance est précise à 1%, cela signifie que l'écart -type entre différentes mesures représente 1% de la valeur pesée. Si l'on pèse plusieurs fois une masse d'un gramme, 95% des pesées seront comprises entre 0,98 g et 1,02 g. Si l'on utilise la même balance pour peser 100 g, 95% des pesées seront comprises entre 98 et 102 g.

Pour une même erreur relative l'erreur absolue est passée de 0,02 g à 2 g!

10.5.7 Centrage

Une mesure centrée est obtenue en retirant la moyenne de l'échantillon à toutes les observations :

$$x_i^{'} = x_i - Mx$$

Par définition, les données centrées ont une moyenne nulle.
Centrons les mesures du tableau ci -dessus :

	Peleds	
	1 semaine	1 an
(X -Mx)i	-1 g	-10 g
	-1 g	-10 g
	1 g	10 g
	1 g	10 g
M(x -Mx)	0 g	0 g
S(x -Mx)	1 g	10 g

Tableau 10-13 Statistiques du poids des peleds et lavarets après centrage : (X -Mx) = variable centre, M(X -Mx) = moyenne nulle, S(X -Mx) = écart -type inchangé.

Suivant les propriétés algébriques de la variance, l'écart -type n'a pas été modifié par l'opération.

10.5.8 Standardisation

Une mesure standardisée est un écart à la moyenne, exprimé en nombre d'écarts -types. On l'appelle généralement Z ou Zscore.

$$Z_i = \frac{x_i - Mx}{S_x}$$ **Équation 10-10**

Standardisons les mesures du tableau ci -dessus :

	Peleds	
	1 semaine	1 an
zi	-1	-1
	-1	-1
	1	1
	1	1
Mz	**0**	**0**
Sz	**1**	**1**

Tableau 10-14 Statistiques du poids des peleds et lavarets après standardisation : zi = variable standardisée, Mz = moyenne nulle, Sz = écart -type unité.

Le Zscore est un nombre sans unité.
Quel que soit le système d'unité de départ, la moyenne est nulle, son écart -type et sa variance valent 1.

10.5.9 Calcul de la variance

En pratique, la SCE utilisée pour le calcul de la variance n'est pas sous la forme de l'équation [1-3]. Elle n'est guère commode, et les imprécisions de calcul sont élevées au carré et sommées, ce qui n'est pas le moyen de calcul le plus précis. Nous pouvons manipuler cette expression afin de la rendre plus efficace:

$$\sum_{i=1}^{n} (x_i - Mx)^2$$

$$\sum_{i=1}^{n} ({x_i}^2 + Mx^2 - 2x_i Mx)$$

$$\sum {x_i}^2 + nMx^2 - 2\sum x_i Mx$$

$$\sum {x_i}^2 + nMx^2 - 2nMxMx$$

$$\sum {x_i}^2 - nMx^2$$

et la variance devient donc :

$$S^2_{(n)} = \frac{SCE}{n} = \frac{\sum x_i^2}{n} - Mx^2 = \frac{30}{4} - \left(\frac{10}{4}\right)^2 = 1.25$$

ou encore

$$\frac{\sum x_i^2 - \frac{(\sum x)^2}{n}}{n} = \frac{30 - \frac{10^2}{4}}{4} = 1.25$$

Exemple :

	x_i	x_i^2
	2	4
	3	9
	4	16
	1	1
Σ	**10**	**30**

Tableau 10-15 Préparation du calcul de la variance : valeurs individuelles, valeurs individuelles au carré, et totaux.

Le passage d'une définition de la variance à l'autre se fait aisément :

$$S^2_{(n-1)} = \frac{SCE}{n-1} = \frac{S^2_{(n)} \times n}{n-1}$$

La plupart des calculatrices électroniques possèdent une fonction qui permet de calculer automatiquement ces variances mais la distinction entre $S^2_{(n)}$ *et* $S^2_{(n-1)}$ *n'est pas toujours évidente, au point que la variance est parfois pré -calculée à partir d'une définition et l'écart -type à partir de l'autre…*

Dans le tableur Excel, l'expression [1-5] correspond à la fonction VAR.P(…) et l'expression [1-6] à la fonction VAR(…).

Dans le tableur Excel, l'expression [1-7] correspond à la fonction ECARTYPEP(…) et l'expression [1-8] à la fonction ECARTYPE (…).

11 Statistiques descriptives

Le premier souci de l'expérimentateur sera de chercher à mettre de l'ordre dans ses observations, de façon à pouvoir les comprendre et les synthétiser pour pouvoir ensuite les comparer et les transmettre.

11.1 Définitions

La population est l'ensemble des observations individuelles limitées dans l'espace et dans le temps, au sujet desquelles on désire tirer des conclusions.

La population doit être définie en fonction de l'objectif de l'étude.

Supposons que je décide de mener une enquête sur l'évolution de la consommation de tabac. Veut -on étendre les conclusions à la population mondiale? Occidentale? Belge? S'intéresse -t -on à toutes les catégories de la population? Seulement aux femmes? Cible -t -on seulement les jeunes? Les moins de 15 ans? Je peux définir que ma population représente les immigrés de 18 à 25 ans, à Bruxelles. Dans ce cas, les conclusions de l'enquête ne pourront absolument pas être étendues à une autre population.

L'échantillon est l'ensemble des observations individuelles sélectionnées dans la population.

*Il est généralement impossible de réunir l'information relative à toutes les observations individuelles comprises dans la population. L'**échantillonnage** est le procédé suivant lequel on sélectionne l'information. Il doit être guidé par la définition de la population. Si par exemple je veux pouvoir parler de la consommation de tabac des hommes en général, mon échantillonnage doit permettre à un Japonais, un Pygmée et un Papou d'avoir les mêmes chances de faire partie de l'échantillon.*

L'échantillon doit être représentatif de la population.

Une variable est la caractéristique numérique mesurée lors de l'observation.

Par exemple, la quantité de nicotine absorbée par le fumeur sera évaluée par la concentration de nicotine dans l'urine. La variable est notée X et les différentes valeurs x_1, x_2, x_3 ... de façon générale ou 6,2, 4,3, 7,8 μg/l d'urine dans un cas particulier, De façon générale, nous dirons avoir réalisé n *observations, et une observation particulière sera notée* x_i, (i = 1,...n).

Une variable discrète est une variable qui montre une discontinuité.

Dans une famille, il y a 1, 2, 3 enfants, pas 3,5. On dira que cette variable peut prendre une infinité dénombrable de valeurs.

Une variable continue est une variable qui peut prendre un nombre infini de valeurs dans un intervalle donné.

Beaucoup de variables naturelles sont des variables continues : le pH d'une solution, le volume d'un organe, la concentration d'une substance dans le sang ... Expérimentalement, une variable est toujours rendue discrète par la précision de la mesure effectuée. D'après la balance que j'utilise, la masse corporelle d'un homme sera de 79 kg, 78,5139 kg, 78,513914 kg ... Elles seront néanmoins considérées comme des variables continues. On dira qu'une variable continue peut prendre une infinité non dénombrable de valeurs.

11.2 Statistiques descriptives à une dimension

Celles -ci se limitent à décrire une seule variable à la fois, qui sera discrète (le nombre d'enfants par famille) ou continue (concentration naturelle des eaux en Cd, on effectue la mesure sur 150 échantillons d'eau en ng/l)...

11.2.1 Tables et graphiques des variables discrètes

Prenons l'exemple de la variable X : nombre d'enfant par famille. J'étudie cette variable dans la population des familles belges, parmi laquelle je sélectionne 50 familles. J'obtiens une série statistique de 50 valeurs, x_1, x_2, x_3 .... x_i, ... x_{50}. Si

je me contente d'énumérer les 50 valeurs, c'est -à -dire de produire la série statistique, l'information ne sera guère pratique.

> *Comment dès lors répondre facilement aux questions : quelle est la proportion de familles de moins de deux enfants, quelle est la proportion de familles bénéficiant d'une réduction au chemin de fer, la proportion de familles de trois enfants est -elle différente en Belgique et en Suède???*

Une façon commode de représenter les résultats consiste à créer une **distribution statistique** des valeurs numériques. La **distribution de fréquence** reprend l'ensemble des k valeurs différentes observées, classées par ordre croissant, x_1, x_2, ... x_i,... , x_k, et les fréquences correspondantes, n_1, n_2, ...,n_i, ..., n_k, la fréquence n_i étant le *nombre de fois* que j'observe une valeur particulière x_i dans l'échantillon.

n étant l'effectif (ici 50), on a évidemment la relation :

$$\sum_{i=1}^{k} n_i = n$$

Équation 11-1

Cette équation implique que chaque observation peut être classée dans une et une seule des k catégories définies.

> *Toute statistique se réfère à des <u>conventions</u> : il faudra décider comment classer une femme enceinte, un enfant décédé, les enfants de familles recomposées...*

En général, on définira toutes les catégories correspondant aux valeurs discrètes comprises entre les valeurs minimum et maximum observées; si certaines valeurs ne sont pas observées, on associera à ces catégories une fréquence nulle.

Nb d'enfants :	Belgique	Suède
0	15	19
1	10	25
2	13	11
3	6	5
4	3	4
5	3	2
6	0	1
	50	67

Tableau 11 -11-1 Distributions de fréquences du nombre d'enfants par famille dans deux pays.

Cela fait, on y voit déjà plus clair dans les résultats. Cependant, si je veux comparer l'échantillon belge à l'échantillon suédois, et que celui -ci comprend

un effectif de 80 observations, le niveau absolu des mesures est un obstacle à la comparaison. 19 n'est pas comparable comme tel à 15, puisqu'il s'agit de 19 parmi 67 et 15 parmi 50.
J'aurai donc intérêt à établir **la distribution de fréquences relatives** dans laquelle chaque fréquence est exprimée en proportion (comprise entre 0 et 1) ou en pourcentage (compris entre 0 et 1) de l'effectif.

$$n'_i = \frac{n_i}{n} \quad ou \quad n'_i = \frac{n_i}{n} \times 100$$ **Équation 11-2**

$$\sum_{i=1}^{n} n'_i = 1 \quad ou \quad \sum_{i=1}^{n} n'_i = 100$$

Enfin, les **distributions de fréquences cumulées** permettent de répondre facilement à des questions du type : *quelle est la proportion de familles profitant d'une réduction au chemin de fer?* La fréquence cumulée est établie en additionnant les fréquences de proche en proche à partir de la première valeur n1. La fréquence relative cumulée est définie de la même façon. Si N'i est la fréquence relative cumulée correspondant à la catégorie i, on peut écrire :

$$N_i = \sum_{j=1}^{i} n_j \quad et \quad N'_i = \sum_{j=1}^{i} n'_j$$ **Équation 11-3**

$N_k = n$, $N'_k = 1$ ou 100

A

X	n_i	n'_i	N_i	N'_i
x = 0	15	0,3	15	0,30
x = 1	10	0,2	25	0,50
x = 2	13	0,26	38	0,76
x = 3	6	0,12	44	0,88
x = 4	3	0,06	47	0,94
x = 5	3	0,06	50	1,00
T	50	1		

B

Nb d'enfants :	Belgique	Suède
0	30%	28%
1	20%	37%
2	26%	16%
3	12%	7%
4	6%	6%
5	6%	3%
6	0%	1%
	n = 50	n = 67

Tableau 11 -11-2 **A** *Distribution de fréquences, fréquences relatives, cumulées, et relatives cumulées pour les données belges.* **B.** *Distributions de fréquences relatives pour les données belges et suédoises, exprimées en pourcentage.*

On peut ainsi calculer immédiatement la proportion de familles qui ont 3 enfants et plus :
$X \geq 3 = 1 - X \leq 2 = (1 - 0{,}76) = 0{,}24$.

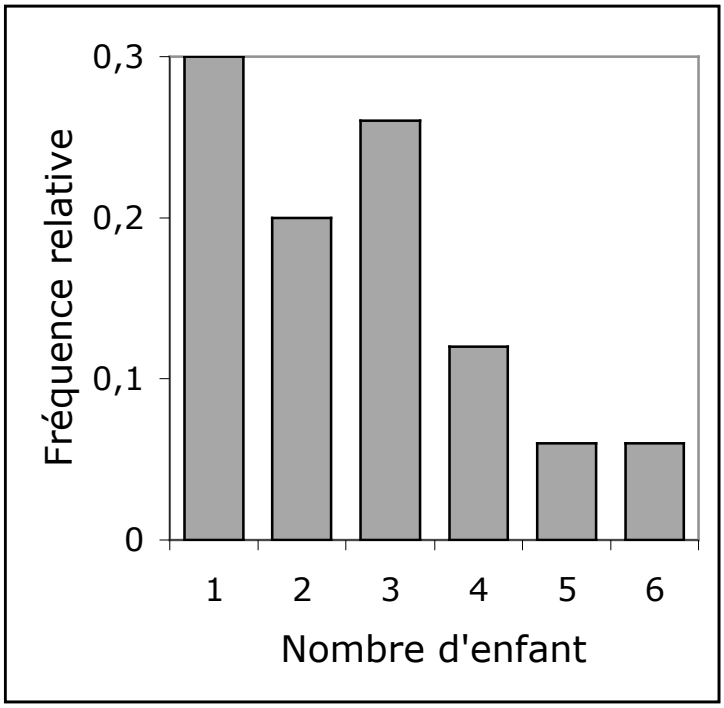

Figure 11 -11-1 Diagramme de barres représentant les fréquences relatives de la table 2.2A.

A ces distributions on associe un graphique, qui se justifie principalement lorsque le nombre de catégories est élevé. (Figure 2.1)

L'abscisse correspond aux différentes valeurs discrètes prises par la variable X, et l'ordonnée représente une des distributions de fréquence.

Un tel diagramme est appelé **diagramme de barres,** en raison de la discontinuité de l'abscisse. Lorsqu'une fréquence relative est exprimée en ordonnée, le nombre total d'observations doit être mentionné sur le graphique.

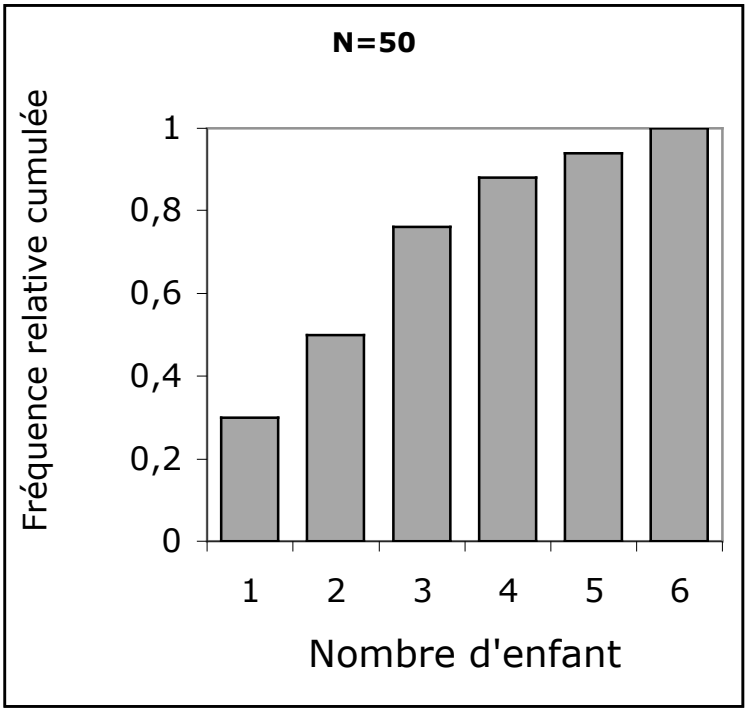

Figure 11 -11-2 Diagramme de barres représentant les fréquences relatives cumulées de la table 2.2A.

11.2.2 Tables et graphiques des variables continues

Les métaux lourds exercent une importante série d'effets sur les poissons. Ceci va des effets métaboliques aux effets physiologiques en passant par des modifications comportementales. Parmi ces métaux lourds, le cadmium est un des éléments les plus communs. Il est souvent déversé avec les effluents industriels et domestiques. Chez le poisson, le Cd a des effets sur la croissance et est responsable de stress osmorégulateur.[2] Il a même été montré que le Cd pouvait altérer la structure et la fonction de divers organes tel que le foie. La réponse à un stress est encore mal comprise et en particulier les effets spécifiques du Cd sur les voies métaboliques de production d'énergie. Ce rapport met en évidence les effets du Cd sur plusieurs voies du métabolisme des hydrates de carbone dans le foie des saumons atlantiques.

Afin de déterminer la concentration naturelle des eaux en Cd, on effectue la mesure sur 150 échantillons d'eau (ng/l). Les valeurs, continues, ont été tronquées à deux chiffres significatifs. (Tableau 2.3)

[2] Osmorégulateur : qui régule l'osmose, mécanisme impliqué dans le maintien d'une concentration ionique intracellulaire différente de celle du milieu.

Les résultats sont relativement encombrants. Il faudrait un long examen rien que pour établir que la concentration observée est comprise entre 33 et 66 ng/l.

57	44	44	48	45	52	49	45	45	52
40	56	60	51	53	56	42	40	63	55
47	49	47	41	49	61	54	44	46	57
54	61	47	60	48	54	52	43	46	43
50	54	50	45	58	51	51	53	53	49
39	54	35	52	52	58	52	41	64	55
49	43	47	49	55	55	48	37	51	52
45	53	66	33	63	47	66	39	43	55
64	66	54	50	51	55	64	34	55	57
47	49	49	33	40	52	52	51	56	54
47	50	42	49	60	47	60	50	50	51
49	41	55	51	46	52	58	58	61	43
41	46	55	49	42	53	50	56	45	57
45	49	46	53	48	41	51	58	46	53
46	42	53	41	62	49	44	42	53	50

Tableau 11 -11-3 Détermination du taux de Cd dans 150 échantillons d'eau (ng/l).

Une première synthèse est obtenue en regroupant d'abord les données en classes.
Ceci implique de définir un intervalle de classe, généralement constant, qui divise l'intervalle (maximum -minimum) en une série d'intervalles plus petits (limite supérieure -limite inférieure). Les données appartenant à cet intervalle sont assignées à la classe correspondante.

A chaque classe est associée une fréquence, qui correspond au nombre d'observations individuelles assignées à cette classe. Les classes doivent toujours être exclusives, de façon à réaliser l'équation 2.1.

Pour cela, il faut classer arbitrairement les valeurs qui correspondent aux limites de classes, soit systématiquement dans la catégorie supérieure ou inférieure, soit alternativement dans l'une et dans l'autre. Ce problème est lié au fait que la variable continue a artificiellement un caractère discret de par la limitation du nombre de chiffres significatifs.

Les fréquences, fréquences relatives et fréquences relatives cumulées peuvent dès lors être définies de la même façon que pour les variables discrètes, chaque classe étant identifiée soit par ses limites, soit par son centre (valeur à équidistance des limites).

limites	centre	ni	Ni	n'i	N'i
31-33	32	2	2	1,3%	1,3%
33-35	34	2	4	1,3%	2,7%
35-37	36	1	5	0,7%	3,3%
37-39	38	2	7	1,3%	4,7%
39-41	40	9	16	6,0%	10,7%
41-43	42	10	26	6,7%	17,3%
43-45	44	11	37	7,3%	24,7%
45-47	46	15	52	10,0%	34,7%
47-49	48	17	69	11,3%	46,0%
49-51	50	17	86	11,3%	57,3%
51-53	52	19	105	12,7%	70,0%
53-55	54	16	121	10,7%	80,7%
55-57	56	8	129	5,3%	86,0%
57-59	58	5	134	3,3%	89,3%
59-61	60	7	141	4,7%	94,0%
61-63	62	3	144	2,0%	96,0%
63-65	64	3	147	2,0%	98,0%
65-67	66	3	150	2,0%	100,0%
total		150		100%	

Tableau 11 -11-4 Tables de fréquences, fréquences relatives, cumulées et relatives cumulées pour 150 échantillons d'eau (Cd, ng/l). La fréquence est calculée en incluant la limite de droite.

Le nombre de classes est arbitraire. On se rend aisément compte qu'il doit réaliser un compromis entre deux extrêmes : une seule classe, ce qui supprime presque toute l'information, et autant de classes qu'il y a de valeurs différentes, ce qui ne réalise aucune synthèse des résultats. Généralement, le nombre de classes est proportionnel au nombre d'observations, l'intervalle de classe est constant, et les classes de fréquence nulle sont évitées.

Ensuite un graphique peut être réalisé, en plaçant :
- en abscisse la variable continue (le centre, ou les limites, ou une sélection de ces valeurs, est indiquée en fonction de l'espace disponible)
- en ordonnée la fréquence, la fréquence relative ou la fréquence relative cumulée. On représente la fréquence par une série de rectangles contigus, ce qui indique le caractère continu de la variable.

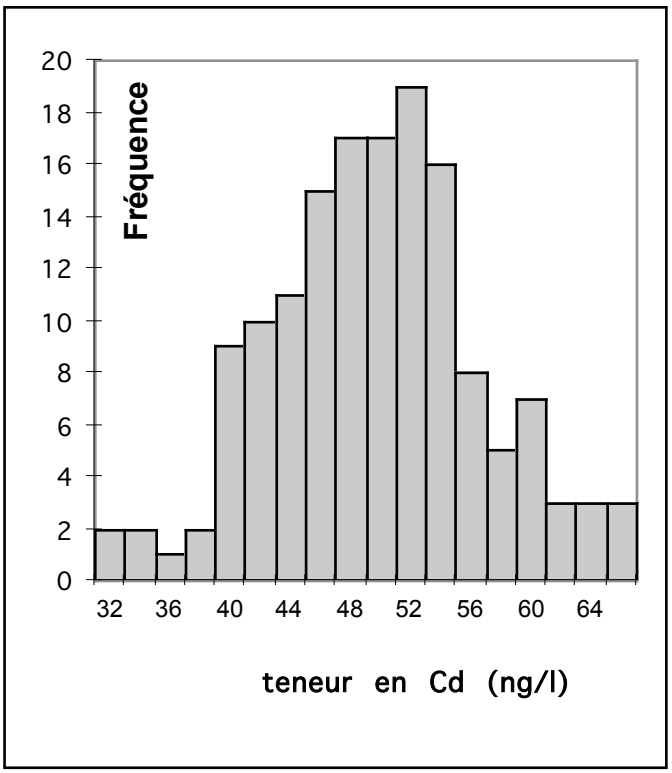

Figure 11 -11-3 Histogramme représentant la distribution de fréquence de la teneur en Cadmium dans 150 échantillons d'eau.

Si l'intervalle de classe Δ est constant, la hauteur du rectangle, mais aussi sa surface, sont proportionnelles à la fréquence relative correspondant à la classe.

Pour comparer des situations dont l'intervalle de classe varie, il est préférable de calculer la densité de fréquence relative qui se définit en divisant la fréquence relative par l'intervalle de classe :

$$d.f.r. = \frac{f.r.}{\Delta}$$ **Équation 11-4**

Si nous comparons un nouvel échantillon de 247 valeurs représenté en 5 ou en 50 classes (figure 2-4), avec en ordonnée la fréquence relative, nous constatons que la surface représentée est beaucoup plus petite lorsque le nombre de classes est élevé.
Nous pouvons constater que si l'on représente en ordonnée la densité de fréquence relative d.f.r., la surface S du rectangle reste proportionnelle à la fréquence relative f.r. quel que soit l'intervalle de classe Δ.

En effet :

$$S = base \times hauteur = d.f.r. \times \Delta = \frac{f.r.}{\Delta} \times \Delta = f.r.$$

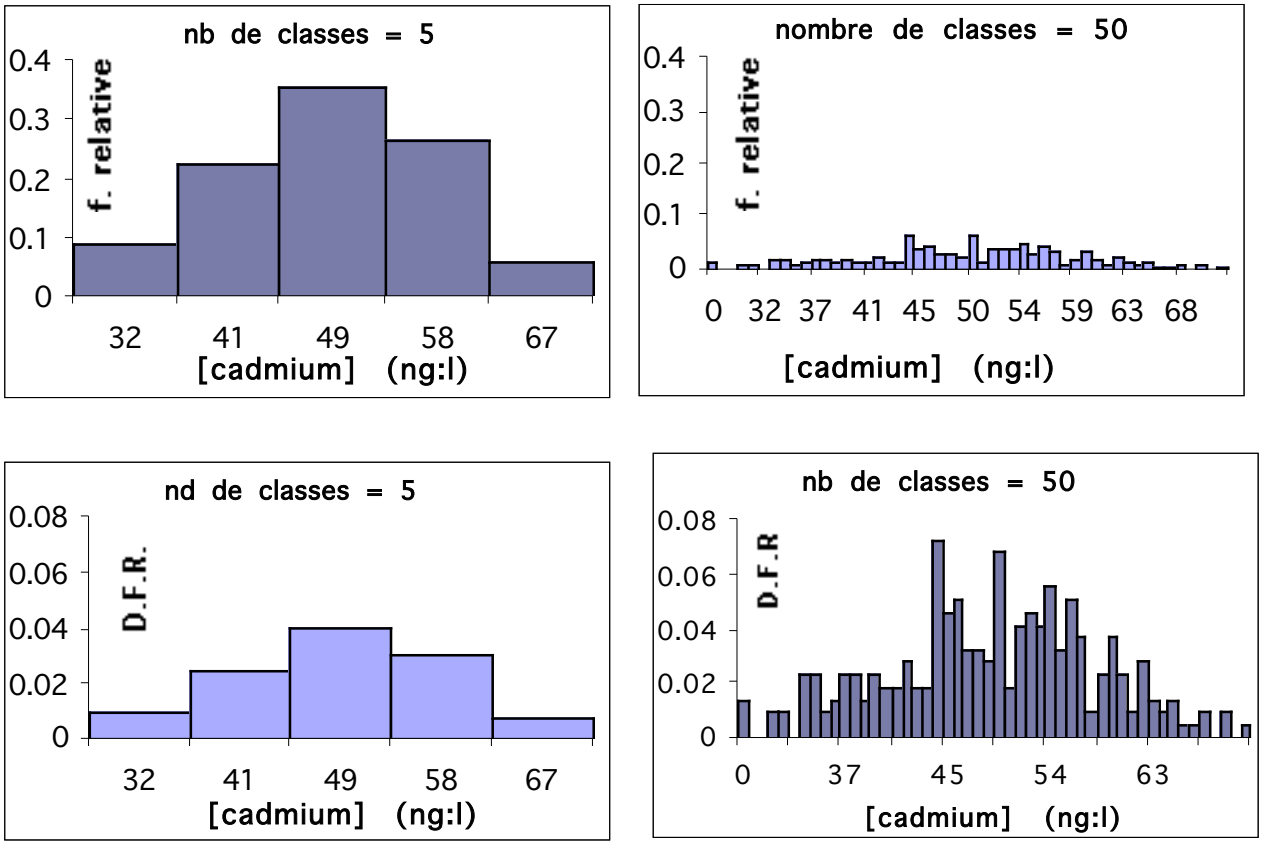

Figure 11 -11-4 Fréquence relative et densité de fréquence relative de la teneur en Cadmium dans 247 échantillons d'eau, pour 5 et 50 classes. L'intervalle de classe plus petit fait ressortir une surface plus faible lorsque l'ordonnée est la fréquence relative.

Ces distributions étant établies et représentées graphiquement, l'expérimentateur est mieux à même d'appréhender l'information qu'il a récoltée. Mais si les tables et les graphiques sont très utiles pour comprendre ces résultats, ils restent encore trop encombrants pour la suite du travail.

On ne dispose souvent que de quelques lignes pour transmettre les résultats d'une expérience donnée. Nous allons donc rechercher le moyen de transmettre un maximum d'informations en un minimum de valeurs numériques.

Imaginons par exemple que je cherche à vous expliquer par téléphone, en un minimum de mots, l'information principale contenue dans le

graphique. Je peux vous dire que le graphique a grosso modo *l'allure d'une pyramide, qu'il est centré sur une valeur de 50 ng/l et que les valeurs sont comprises entre 33 et 66 ng/l.*

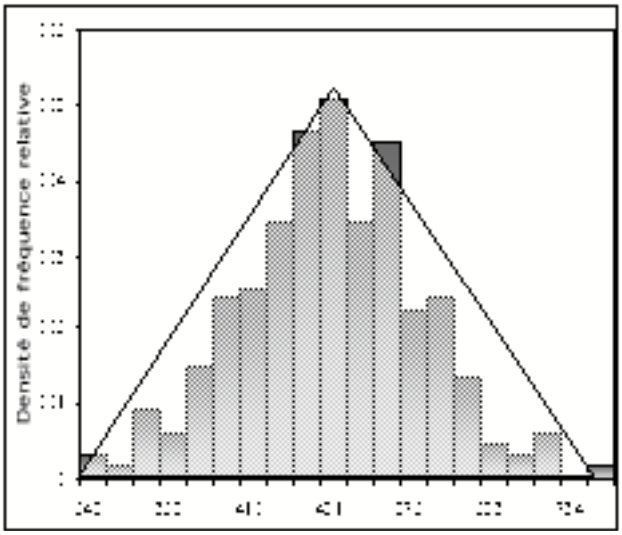

Figure 11 -11-5 Représentation schématique de la teneur en Cadmium dans 150 échantillons d'eau. Compte tenu d'une distribution symétrique, l'essentiel de l'information peut être synthétisé par une mesure de tendance centrale et une mesure de dispersion.

Je dois donc essentiellement communiquer une valeur exprimant la tendance centrale et une valeur exprimant la dispersion de la distribution.

11.2.3 Tendance centrale d'une distribution

Définir le centre d'une distribution, c'est trouver un nombre caractéristique de sa position. Il existe pour cela plusieurs possibilités.

❖ La moyenne arithmétique

La moyenne arithmétique est certainement le moyen le plus courant d'exprimer la tendance centrale d'une distribution. Pour n observations x_1, x_2, ..., x_i, ..., x_n, elle est définie comme :

$$Mx = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \{2.6\}$$

Si l'on dispose seulement des valeurs regroupées en k classes, on définit x_1, x_2, ..., x_i, ... x_k comme étant les centres de classes, et en considérant n_i la fréquence de chaque classe, on peut estimer

$$Mx = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \quad \{2.7\}$$

La moyenne arithmétique est simple à calculer, et elle présente l'avantage algébrique d'assurer que la somme des écarts à la moyenne est nulle et que la somme des carrés des écarts peut être décomposée en composantes additives
(voir chapitre 1).

La moyenne arithmétique présente un inconvénient important : lorsque la distribution n'est pas symétrique, elle ne représente pas bien la tendance centrale de la distribution.

Influencée par les valeurs extrêmes, elle peut être nettement (sous -) surestimée. D'autres mesures peuvent être intéressantes à considérer en cas d'asymétrie de la distribution : le mode et la médiane.

❖ **Le mode**

Le mode est le centre de la classe pour laquelle la fréquence est la plus élevée. C'est une valeur à laquelle on fait facilement référence intuitivement : on dira par exemple que la famille type dans nos sociétés est de deux enfants, car c'est la catégorie de familles la plus fréquente. Pourtant, la moyenne arithmétique est de 1.7 enfants par couple. Le mode présente plusieurs inconvénients, il dépend d'un intervalle de classe arbitraire, il est sensible à la taille de l'échantillon.

❖ **La médiane**

est la valeur au -dessous de laquelle se trouve la moitié des observations. C'est donc une valeur qui coupe l'échantillon en deux parties égales. Cette mesure peut parfois donner une information fort intéressante.

Ceci illustre que si la distribution de la variable est symétrique, la moyenne arithmétique, le mode et la médiane sont confondus.

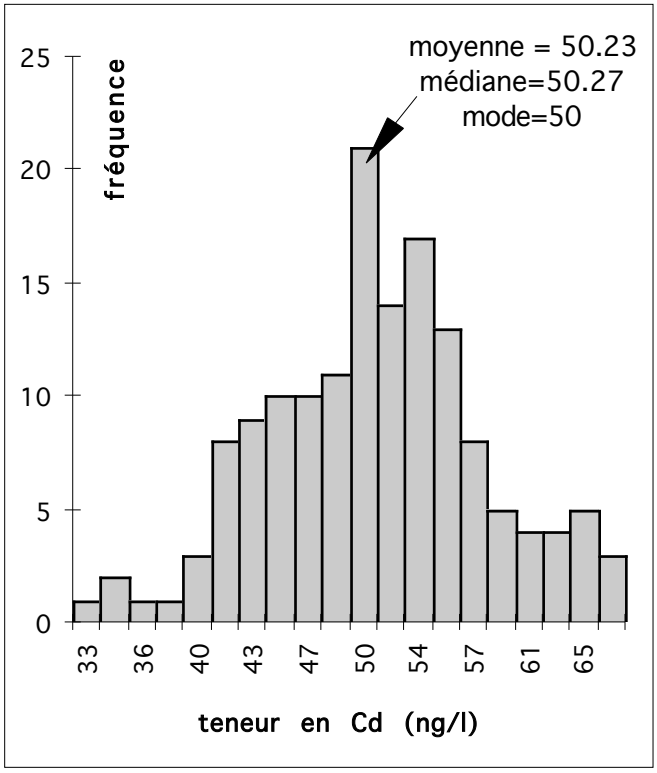

Figure 11 -11-6 Moyenne, médiane et mode dans une distribution symétrique. Les valeurs de représentation de la tendance centrale sont très similaires dans une distribution symétrique.

Si la distribution est asymétrique (nombre d'enfants par famille, salaire mensuel) le mode et la médiane donnent des informations différentes de la moyenne arithmétique, dont le principal inconvénient est d'être très influencée par quelques valeurs extrêmes .

Supposons (Tableau 2.5) que l'on puisse dire que dans deux pays la médiane du revenu est de 2 000 Euros par mois. En calculant la moyenne, on peut mettre en évidence que dans le premier pays, la moyenne est de 2 765 Euros par mois et que dans l'autre où quelques-uns atteignent des revenus beaucoup plus élevés, la moyenne est de 4 280 Euros par mois !

Revenu	Pays		Pays	
Euros/ mois	A		B	
	n'i (%)	Ni (%)	n'i (%)	Ni (%)
500	10	10	12	12
2000	40	**50**	38	**50**
3500	39	89	33	83
5000	11	100	9	92
15500	0	100	1	93
18500	0	100	2	95
24500	0	100	2	97
26000	0	100	1	98
29000	0	100	2	100
Total	100		100	
Moy.	2765		4280	

Tableau 11 -11-5 Revenu moyen hypothétique dans deux pays. Calcul des distributions de fréquences relatives (%) et de fréquences relatives cumulées (%) sur 1.9 millions de personnes. Calcul de la moyenne et de la médiane.

Dans le tableau 2-5, il apparaît clairement que la moyenne arithmétique surestime la tendance centrale lorsque la distribution est asymétrique : les rares salaires très élevés "attirent" la moyenne vers le haut.

En conséquence, lorsque la distribution est asymétrique, la définition de la tendance centrale est équivoque et les différents indicateurs produisent des informations différentes.

11.2.4 Dispersion d'une distribution

Supposons que vous lisiez dans un article médical que pour un échantillon de 1000 individus normaux le taux de cholestérol dans le sang présente une moyenne de 154,5 mg/dl de sang. Si vous obtenez un résultat d'analyse vous concernant signifiant que votre taux de cholestérol est de 170,3 mg/dl, vous allez éventuellement vous alarmer : "je ne suis pas normal !*", Cependant, la question qu'il faut préalablement se poser est la suivante :* Etant donné la variabilité individuelle du taux de cholestérol, est -il improbable qu'un individu normal (admettons une population de moyenne 155 mg/dl) présente un taux de 170 mg/dl?

C'est -à -dire que vous vous demandez si les valeurs individuelles peuvent être sensiblement éloignées de la valeur moyenne. En fait, la revue médicale pourrait vous fournir directement ce renseignement, sous la forme d'une mesure de dispersion.

Une moyenne ne constitue pas en elle -même une information opérationnelle, susceptible de guider une interprétation.

❖ L'amplitude

est une mesure de dispersion à laquelle on fait immédiatement référence intuitivement. Pour déterminer l'amplitude, on recherche la valeur minimum (ex : 126,6 mg/dl) et la valeur maximum (ex : 181,2 mg/dl) observées dans l'échantillon, et on les soustrait, ce qui donne une amplitude de 54,6 mg/dl. Le principal inconvénient de cette notion simple est qu'elle est extrêmement sensible à la taille de l'échantillon. En effet, on observe immédiatement sur l'histogramme qui représente la distribution de fréquence que les valeurs les plus éloignées de la moyenne sont les moins fréquentes. Par conséquent, plus l'échantillon sera petit, moins on aura de chances d'observer ces valeurs extrêmes, ce qui va changer la mesure d'amplitude. L'amplitude sera donc en partie le simple reflet de la taille de l'échantillon.

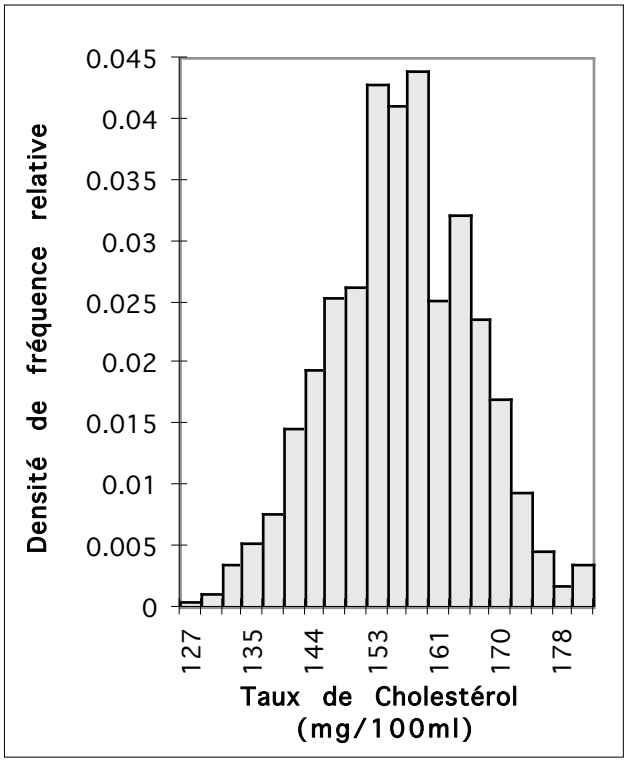

Figure 11 -11-7 Distribution de 1000 observations du taux de cholestérol.

❖ L'écart -type

La valeur la plus couramment utilisée est l'écart -type noté S (ou S_x, pour indiquer qu'il s'agit de la variable X).

Dans cet échantillon, l'écart -type est de 9,83 mg/dl de sang.

Il faut bien remarquer que toutes les valeurs ne sont pas comprises entre la moyenne ± l'écart -type, puisqu'il s'agit d'un écart moyen. Si la distribution est symétrique, on observe approximativement :

environ 68% des valeurs comprises entre la moyenne ± 1 écart –type ;

environ 95% des valeurs comprises entre la moyenne ± 2 écarts –types ;

environ 99% des valeurs comprises entre la moyenne ± 3 écarts –types.

L'interprétation la plus courante de l'écart -type est liée à l'intervalle à 95% :

Dans une distribution symétrique, l'écart -type est une valeur telle que 95% des valeurs environ sont comprises entre la moyenne ± 2 écarts –types.

Vous pouvez maintenant juger, compte tenu de cette information, que 95% des individus ayant un taux de cholestérol sanguin compris entre 134,8 et 174,2 mg/100 ml, votre taux de 170,3 révèle une concentration élevée sans être anormale[3].

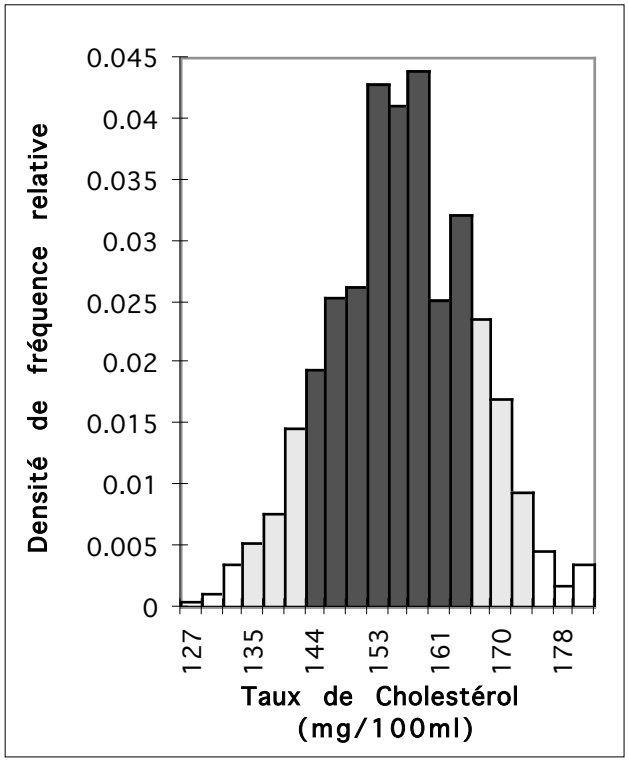

Figure 11 -11-8 Distribution de 1000 observations du taux de cholestérol dans une population d'individus normaux : en foncé, les observations comprises entre la moyenne ± 1 écart -type, en gris, les observations comprises entre la moyenne ± 2 écarts -types.

[3] *Attention : une situation médicale « normale », c-à-d non pathologique, ne doit pas être confondue avec une variable normale, c-à-d ayant des propriétés statistiques particulières.*

		inf	sup
Potassium	mmol/l	3,1	4,9
Glucose	g/l	0,6	1
Urée	g/l	0,15	0,41
Créatinine	mg/l	6	11
Acide urique	mg/l	30	65
Protéines totales	g/l	60	80

Tableau 11 -11-6 Extrait d'un protocole d'analyse sanguine : chaque variable est accompagnée des limites à 95% des valeurs normales,

11.2.5 Représentation graphique de la moyenne et de l'écart -type

Une technique graphique similaire à celle du diagramme de barre peut être utilisée pour représenter la moyenne de différents échantillons. On l'accompagne souvent d'une représentation de la variabilité des observations (par exemple ± 2S).

Un systématicien s'intéresse à un poisson Cyprinidae du genre Labeo afin de déterminer si les spécimens capturés dans deux rivières du Congo (rivière Kouilou et Shiloang) appartiennent à la même espèce ou à deux espèces différentes. Les résultats sont représentés dans les figures suivantes avec et sans indication de l'écart -type.

Notez que les moyennes apparaissent très semblables dans le premier graphique, et très différentes dans le second, suite à l'ajustement de l'échelle à la variabilité individuelle.

Notez que les barres d'erreur peuvent représenter, suivant le choix de l'auteur, un écart -type, deux écarts -types, un pourcentage (ou encore une ou deux erreurs standards, notion qui sera développée plus loin). Il faut donc être prudent lors de leur interprétation, surtout qu'elles ne sont pas nécessairement définies clairement (par négligence de l'auteur).

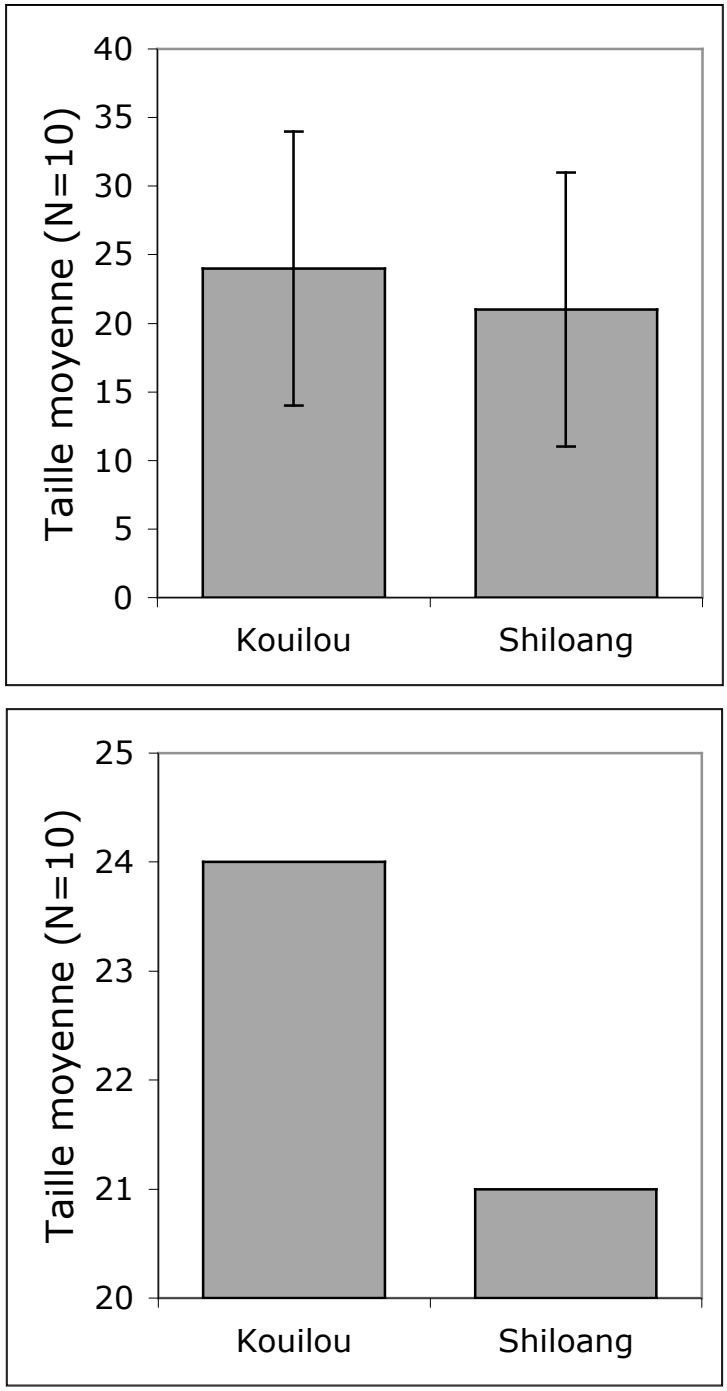

Figure 11 -11-9 Représentation de la moyenne ± 2S de deux échantillons.