INSTITUT SUPERIEUR DE L'EDUCATION

ET DE LA FORMATION CONTINUE

Département Biologie – Géologie

SN101/1

" BIOSTATISTIQUE - 1 "

Cours & Activités : Mondher Abrougui

Année Universitaire - 2008

PLAN DU COURS - BIOSTATISTIQUE

BIOSTATISTIQUE 1 = L1 : Statistiques descriptives à une ou deux variables
BIOSTATISTIQUE 2 = L2 : Statistiques inférentielles à une ou deux variables
BIOSTATISTIQUE 3 = L3 : Statistiques multifactorielles descriptive et inférentielles

BIOSTATISTIQUE 1
STATISTIQUES DESCRIPTIVES À UNE OU DEUX VARIABLES

CHAPITRE I.
ELEMENTS DE STATISTIQUE ET DE BIOSTATISTIQUE

1. INTRODUCTION A LA STATISTIQUE

2. OBJECTIF DES MODULES DE BIOSTATIQUE
2.1. Le module biostatistique I : Statistiques descriptives
2.2. Le module biostatistique II : Statistiques inférentielles
2.3. Le module biostatistique III : Statistiques multifactorielles descriptive et inférentielle

3. DEMARCHE GENERALE EN STATISTIQUE
3.1. L'identification du problème
3.2. Le recueil des données
3.3. L'analyse et l'interprétation des données

4. NOTIONS DE BASE ET TERMINOLOGIE
4.1. Ensemble / Population / Echantillon / Elément / Individu
4.2. Recensement / Echantillonnage
4.2.1. Le recensement
4.2.2. L'échantillonnage
4.3. Caractère / Modalité / Variable:
4.3.1. Le caractère
4.3.2. Modalité / Mesure
4.3.3. Tableau élémentaire
4.3.4. La variable statistique
4.3.5. Nature des variables statistiques et échelles de mesures
4.3.5.1. Variable quantitatif.
4.3.5.2. Variable qualitative
4.3.5.3. Exemple d'illustration des principaux types de descripteurs
4.3.6. Variables dépendantes et indépendantes
4.3.6.1. Les variables indépendantes
4.3.6.2. Les variables dépendantes
4.3.7. La variabilité et l'incertain en biologie
4.3.7.1 La variabilité biologique
4.3.7.2 La variabilité métrologique
4.3.8. Propriétés des variables
4.4. Inférence et risque statistique
4.5. Une définition plus explicite de la biostatistique :
4.6. Dénomination mathématique :

5. REPRESENTATION DES DONNEES

5.1. Tableaux statistiques
5.2. Représentations graphiques et statistique descriptive
5.2.1. L'histogramme
5.2.1.1. L'histogramme : paramètres de description (mode et symétrie)
5.2.2. Barre à moustache - Box Plot

6. STATISTIQUES DESCRIPTIVES UNIVARIEES

6.1. Paramètre de position et valeurs centrales
6.1.1. Le mode, ou valeur dominante
6.1.2. La moyenne
6.1.2. 1. Calcul de la moyenne par changement d'origine et d'unité.
6.1.2.2. Autres indicateurs de moyenne :
6.1.3. La médiane et la classe médiane
6.1.3.1. Définition générale :
6.1.3.2. Médiane, pour les données rangées
6.1.3.3. Médiane, pour les données condensées
6.1.3.4. Médiane, pour les données réparties par classes
6.1.4 Quantiles : Mesures de position statistique en référence à la médiane
6.1.4.1. Définition des quantiles
6.1.4.2. Les quartiles
6.1.4.3. Les déciles
6.1.4.4. Les centiles
6.1.4.5. Calculs des quantiles
6.1.4.5.1. Détermination des valeurs de la variable à partir d'un rang centile données.
6.1.4.5.2. Détermination du rang centile à partir d'une valeur donnée de la variable.
6.1.5. Moyenne et médiane
6.1.6. Avantages et inconvénients des différentes valeurs centrales :
6.2. Paramètre de dispersion
6.2.1. Les paramètres de dispersion absolue
6.2.1.1 L'étendue de la variation
6.2.1.2. Quartile et intervalle interquartile : Mesures de la dispersion statistique en référence à la médiane
6.2.1.2.1. L'intervalle interquartile
6.2.1.2.2. L'intervalle inter-décile
6.2.1.3. Mesures de la dispersion statistique en utilisant l'écart semi-interquartile
6.2.1.4. Mesures de la dispersion statistique en référence à la moyenne arithmétique
6.2.1.4.1. Ecart absolu moyen ou Ecart Moyen Absolu « EMA »
6.2.1.2.2. Variance et écart-type :
6.2.2. Les paramètres de dispersion relative
6.3 Exercices d'Applications avec explicitation et utilisation du logiciel Excel
6.3. Paramètres de forme
6.3.1. Coefficient d'asymétrie et de dérive
6.3.1.1. Coefficient d'asymétrie
6.3.1.2. Coefficient de dérive
6.3.2. Coefficient d'aplatissement

PARTIE EXERCICES

CHAPITRE II

ETUDE DE DEUX VARIABLES STATISTIQUES - SERIE STATISTIQUE DOUBLE -

CHAPITRE I.

ELEMENTS DE STATISTIQUE ET DE BIOSTATISTIQUE

1. INTRODUCTION A LA STATISTIQUE

Statistique : le terme statistique désigne à la fois :
1) l'ensemble des données numériques concernant une catégorie de faits (sens très ancien). Il s'agit de l'expression dans sa signification la plus usuelle.
2) l'ensemble des méthodes mathématiques permettant :
a) de résumer quantitativement l'information recueillie sur un ensemble d'éléments au moyen d'une investigation exhaustive. C'est la **statistique descriptive**, qui fait l'objet de ce cours.
b) de généraliser à de grands ensembles d'éléments les conclusions tirées des résultats obtenus avec des ensembles beaucoup plus restreints appelés échantillons. C'est la **statistique inférentielle** ou probabiliste, qui sera brièvement explicitée dans ce module et plus explicitée dans les modules de Biostatistique II et III.

Les statistiques ont pour origine le besoin des États pour gérer rationnellement leurs ressources. Pour cela, il était nécessaire après collecte d'informations (nécessité de techniques de quantification ; production de données nombreuses, organisées en tableaux) de disposer de méthodes permettant de définir les variations, les évolutions, les ressemblances ou les différences entre régions, entre années, entre catégories.

Exemple de problèmes :
Dénombrement des populations humaines : recensements
Dénombrement des terres et leur répartition.
Calcul et répartition des impôts.

Ces techniques se sont mises en place grâce au développement du calcul des probabilités au $18^{ème}$ siècle; puis, au $19^{ème}$ siècle grâce à l'émergence des méthodes statistiques. Il s'agissait au départ de l'étude méthodique des faits sociaux par des procédés numériques : classements, dénombrements, inventaires chiffrés, recensements, destinés à renseigner et à aider les gouvernements dans leurs prises de décisions.

À partir de 1843, la statistique désigne l'ensemble de techniques d'interprétation mathématique appliquées à des phénomènes pour lesquels une étude exhaustive de tous les facteurs est impossible, à cause de leur grand nombre ou de leur complexité. Les statistiques s'appuient sur les probabilités et sur la loi des grands nombres.

La statistique vise à décrire, à résumer et à interpréter des phénomènes dont le caractère essentiel est la variabilité. Elle fournit de la manière la plus rigoureuse possible des éléments d'appréciation utiles à l'explication ou à la prévision de ces phénomènes, mais elle n'explique ni ne prévoit aucun d'entre eux (Vigneron 1997). La méthode statistique permet également d'éprouver la validité de résultats (obtenus, mesurés, collectés) en fonction même de leur variabilité, dans les domaines où les variations sont la règle, c'est-à-dire les domaines de la biologie *sensu lato*, dans celui des sciences de l'environnement également. La méthode statistique fournit de ce fait à tous les personnels confrontés à l'interprétation de résultats d'observation ou d'expérimentation, un outil d'interprétation adapté aux conditions particulières de leur domaine d'activité.

L'attrait des chiffres tient dans la croyance que la mensuration est le critère primordial de toute étude scientifique (Francis Galton (1822-1911)). Cette fascination répond à l'idée que ce que nous voyons et mesurons dans le monde n'est que la représentation superficielle et imparfaite d'une réalité cachée. Il faut se méfier de cette tendance qui veut que les mesures abstraites résumant de grands tableaux de données doivent exprimer nécessairement quelque chose de plus réel et de plus fondamental que les données elles-mêmes. Tout statisticien doit faire un effort pour contrebalancer cette tendance. C'est pourquoi toute interprétation statistique doit préciser clairement quelles données (population, échantillon), et quelles hypothèses ont été utilisées pour aboutir à un certain type de conclusion.

2. OBJECTIF DES MODULES DE BIOSTATIQUE

L'enseignement de la biostatistique est subdivisé en 3 modules : Biostatistique I, Biostatistique II et Biostatistique III. Ces trois modules complémentaires ont pour objectifs de permettre aux étudiants de développer des compétences qui leur permettront :
- d'acquérir et de parfaire la connaissance des principales notions relatives à l'utilisation des méthodes statistiques,
- de résoudre des questions empiriques par l'utilisation des tests statistiques,
- de maîtriser et de compléter les notions de bases des statistiques en vue de les appliquer à des exemples spécifiques aux sciences biologiques, prises dans leur sens général (biologie, médecine, pharmacie, écologie…)
- d'appliquer ces notions et méthodes sur des données biologiques à partir de logiciels simples
- d'utiliser des logiciels de statistique et d'apprendre la lecture de leurs résultats.

Les statistiques constituent, en biologie, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence aux biologistes, en voici, à titre d'exemples quelques unes :
- Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?
- Quelle est la fiabilité d'une mesure ou d'une observation ?
- Quel est le risque ou l'avantage d'un traitement?
- Les conditions expérimentales A sont-elles plus efficaces que celles des conditions de B ?
- Les effets de la variable A sont-ils les mêmes ou diffèrent-ils des effets de la variable B ?

Ces cours visent à développer la compréhension conceptuelle des biostatistiques, à travers l'application, les suppositions sous-jacentes, et l'interprétation d'analyses statistiques présentées avec un minimum de formules et avec l'assistance d'interface et de logiciels informatiques.

2.1. Le module biostatistique I : Statistiques descriptives

Ce module est une initiation aux notions fondamentales de statistique descriptive (non paramétrique et paramétrique). Il explicitera les procédés classiques de la statistique à une dimension, uni-modale, bimodale et uni-variée qui permettent de résumer et d'analyser l'information recueillie sur chaque caractère (variable (continue ou discrète, qualitative ou quantitative)) pris isolément. Ce module de Biostatistique 1, vise à initier les étudiants aux statistiques et à présenter brièvement la première étape de l'analyse des données : **la description.** L'objectif poursuivi dans une telle analyse est de 3 ordres :
- tout d'abord, obtenir un contrôle des données et éliminer les données aberrantes,
- ensuite, résumer les données (opération de réduction) sous forme graphique ou numérique,
- enfin, étudier les particularités de ces données

Ce qui permettra éventuellement de choisir des méthodes plus complexes.
Les méthodes descriptives se classent en deux catégories qui souvent sont complémentaires : la **description numérique** et la **description graphique**.

La présentation synthétique d'un grand ensemble de données résultant de l'étude de plusieurs caractères quantitatifs ou qualitatifs sur une population sera traitée par le module de biostatistique III.

2.2. Le module biostatistique II : Statistiques inférentielles

Ce module reprend les éléments de bases des statistiques descriptives en y introduisant une approche plus probabiliste. Les méthodes statistiques sont orientées vers des études classiques d'estimation et d'hypothèse, de manière à satisfaire les conditions d'applications des méthodes de l'inférence (approche déductiviste). Il fourni des outils statistique qui permettent d'étendre ou de généraliser, dans certaines conditions, les conclusions obtenues par la statistique descriptive à partir de la fraction des individus (échantillon) que l'on a observé ou étudié expérimentalement, à l'ensemble des individus constituant la population. L'objectif de ce module de statistique inférentielle est de fournir des résultats relatifs à une population à partir de mesures statistiques réalisées sur des échantillons ou de comparer statistiquement et de façon significative si des échantillons sont identiques ou non selon un ou plusieurs paramètres ou tests (indépendance, hypothèses, estimation,…).

2.3. Le module biostatistique III : **Statistiques multifactorielles descriptive et inférentielle**

Ce module complète les modules de biostatistique I et II. Il sera centré sur l'étude multifactorielle qui fournit des méthodes visant à décrire l'**information globale** dont on dispose quand on considère les caractères étudiés dans leur ensemble. Les interrelations entre les caractères et leurs effets sur la structuration de la population seront pris en considérations. L'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC) ont pour but de révéler ces interrelations entre caractères et de proposer une structure de la population. Un des intérêts majeurs de ces analyses est de fournir une méthode de représentation d'une population décrite par un ensemble de caractères dont les modalités sont quantitatives (mesures continues), pour une ACP, ou qualitatives (pour une AFC).

3. DEMARCHE GENERALE EN STATISTIQUE

Toute étude statistique peut être décomposée en deux phases au moins : le recueil ou la collecte des données statistiques, et leur analyse ou leur interprétation.

3.1. L'identification du problème

La phase préliminaire à toute approche statistique vise à déterminer et identifier le problème par un ensemble de questionnements qui permettront de délimiter les investigations et les différentes approches :
Quels sont les objectifs ?
Quelle est la population ou l'échantillon à étudier ?
Quels sont les caractéristiques et les variables ?
Que pourra apporter une étude statistique ?

3.2. Le recueil des données

Nous appellerons données les valeurs obtenus et référencées suite à une investigation ou une étude réalisée (mesures, observations, enquêtes,…).
Le recueil des données peut être réalisé soit par la simple observation des phénomènes, soit par l'expérimentation, c'est-à-dire en provoquant volontairement l'apparition de certains phénomènes contrôlés.
Exemple : le rôle de quelques substances (N, P, K) dans la production de biomasse chez les végétaux.

Lorsque les données sont très nombreuses, ou particulièrement difficiles à obtenir, il sera nécessaire pour la mise en oeuvre rationnelle du recueil de définir des méthodes appropriées de collecte. Il s'agira de plans d'échantillonnage ou de plans d'expérience dont la mise en oeuvre sera fonction du type de problème que l'on est amené à résoudre.

Exemple : la numération des mammifères d'une aire protégée : inventaire et recensement.
Il existe de ce fait plusieurs méthodes de collecte des données (voir notions de base et terminologie):

3.3. L'analyse et l'interprétation des données

L'analyse statistique se subdivise en deux étapes :

- La statistique déductive ou descriptive : elle a pour but de résumer et de présenter les données observées sous la forme la plus accessible (simplification et réduction des données, à la fois visuelle et conceptuelle).

- L'analyse inductive ou inférence statistique est l'ensemble des méthodes permettant de formuler en termes probabilistes un jugement sur une population, à partir des résultats observés sur un échantillon extrait au hasard de cette population. Les méthodes statistiques les plus classiques sont celles de l'estimation (estimation par domaine de confiance) et celles de l'épreuve d'hypothèse. Leurs conceptions de base sont dues essentiellement à R.A. Fisher (1890 - 1962). Elle permet d'étendre ou de généraliser, dans certaines conditions, les conclusions obtenues par la statistique descriptive à partir de la fraction des individus (échantillon) que l'on a observé ou étudié expérimentalement, à l'ensemble des individus constituant la population.

Les conditions (de validité) sont liées aux hypothèses faites sur la population contenant les individus et sur la façon dont ont été prises les mesures. Cette phase inductive comporte des risques d'erreur qu'il convient d'apprécier.

Ces deux étapes sont interdépendantes. En particulier, l'observation et l'expérimentation doivent être organisées (protocole) de manière à satisfaire les conditions d'applications des méthodes de l'inférence. L'objectif de la statistique inférentielle est de fournir des résultats relatifs à une population à partir de mesures statistiques réalisées sur des échantillons (prévision, décision..)

4. NOTIONS DE BASE ET TERMINOLOGIE

4.1. Ensemble / Population / Echantillon / Elément / Individu

- L'**ensemble** en statistique, est la collection (finie ou infinie) d'unités, ou d'éléments, sur laquelle porte l'observation. Pour que cet ensemble soit correctement défini, il faut lui donner une définition précise de façon à ce que deux personnes différentes aboutissent toujours à la même liste d'éléments. L'ensemble des éléments observés sera appelé **E**.

- Les **éléments** sont les objets constitutifs de l'ensemble. Ce sont des objets déterminés dont l'appartenance à tel ou tel ensemble E est sans ambiguïté. Les éléments peuvent être désignés par leur position dans le tableau de données : **1** pour le premier, **i** pour un élément quelconque, **n** pour le dernier élément, **N** pour la somme des éléments constituant l'ensemble.

*** Exemple** :
Élément : membre d'une population statistique (spécimen, prélèvement d'eau, individu...)
*** Question**
Quel est l'élément ? Il faut le définir de manière à pouvoir le reconnaître sans ambiguïté.

- La **population** correspond à l'ensemble des **individus** sur lequel porte l'étude ou la prévision, (il est généralement difficile de l'étudier dans sa totalité), et l'**échantillon** représente la fraction de cette population qui est réellement observée ou étudiée :

- Population-cible : ensemble des éléments visés, en principe, par l'échantillonnage.
*** Question**
Quelle est la population-cible ? Il s'agit là de la population sur laquelle on aimerait bien que les conclusions de l'étude portent.

- **Population statistique** : ensemble des éléments effectivement représentés par l'échantillonnage. Les éléments qui la composent se caractérisent par au moins une caractéristique commune et exclusive qui permet de les distinguer sans ambiguïté.
*** Question**
Quelle est la population statistique ? Il faut mentionner la ou les caractéristiques qui permettent de la distinguer de tout autre population statistique.

- **Population biologique**: ensemble des individus d'une même espèce habitant un lieu donné à un moment donné. Notion qui relève davantage de la biologie que de la statistique.
*** Question**
Quelle est la population biologique ? Il faut spécifier le temps et le lieu.

- **Communauté** : ensemble des individus de diverses espèces retrouvés dans un espace et un temps donnés. Notion qui relève davantage de la biologie que de la statistique.
- *Quelle est la communauté ? Il faut spécifier le temps et le lieu.*

Exemples généraux:
- Pour les instituts de sondage, la population étudiée sera un ensemble d'hommes et de femmes occupant une portion définie de l'espace (pays, région, commune) et l'échantillon " représentatif" sera un nombre limité mais représentatif des catégories pertinentes en fonction

du problème posé (âge, sexe, catégories socio-professionnelles, origine géographique, etc.) (Pour la Tunisie, échantillons de 1000 à 1200 individus pour une population de près de 10 millions d'habitants).
- Toute l'eau qui s'écoule d'une rivière à un moment donné constitue la population. Les 20 prélèvements de 10 cm^3 que l'on va analyser constituent l'échantillon.
- Le sang d'une personne peut être considéré comme une population, une prise de sang comme un prélèvement (individu, observation) et l'ensemble des prélèvements sera considéré comme un échantillon.

La notion d'individu est très large : les éléments d'un échantillon ou d'une population sont appelés généralement des individus, cependant cette notion peut être remplacé par plusieurs dénominations: unité statistique, sujet, objet, élément, observation, mesure, doses,… toutefois, dès que la dénomination est choisi aucune ambiguïté ne doit persistée.

4.2. Recensement / Echantillonnage

4.2.1. Le recensement : qui consiste généralement en un recueil d'informations auprès de tous les individus d'une population (ce qui est très difficile dans le cas de la Biostatistique, mais plus facile dans des études démographique). Il est plus adapté à l'étude des populations. Il consiste en un dénombrement de toutes les personnes ou individus ou attributs d'une population dans sa totalité. Il s'agit de la source de données la plus complète dont on dispose sur la population. La méthode est très fastidieuse car rien n'est négligé pour tenir compte de chaque individu. En effet, le recensement est très important puisqu'il s'agit de la seule enquête permettant de brosser un tableau détaillé de toute la population. L'enquête ou la prise de données ou le référencement des attributs couvre toute la population, ce qui facilite la comparaison des renseignements enregistrés.

Exemples : population d'un pays ; pollution mondiale ; animaux en voie de disparition ; génome humain ; ….

4.2.2. L'échantillonnage : qui consiste généralement en un recueil d'informations auprès de quelques individus ou partie d'une population **« l'échantillon »**, (ce qui est généralement le cas en Biostatistique). Parfois l'échantillonnage se fait par sondage (cas en géologie (tremblement de terre), en médecine)

• **Échantillon** : fragment d'un ensemble prélevé pour juger de cet ensemble. Fraction de la population statistique sur laquelle des mesures sont faites pour connaître les propriétés de cette population.
*** Question**
- *quel est l'échantillon ? Quel est son effectif ?*

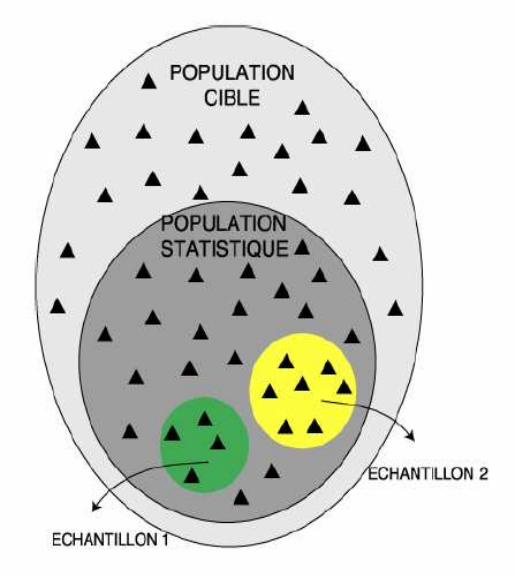


***Figure 1** : Populations et échantillons*

• **Échantillon représentatif** : échantillon qui représente fidèlement la composition et la complexité de la population statistique.
- *pour être représentatif, un échantillon doit être aléatoire.*
- *représentatif ne veut pas dire "conforme à l'idée que le chercheur se fait de la population"!*
• **Échantillon aléatoire**: échantillon prélevé de manière à ce que chacun des éléments de la population statistique ait une probabilité connue et non nulle d'appartenir à cet échantillon.
- *un échantillon aléatoire est représentatif de sa population statistique*
- *aléatoire ne signifie pas "n'importe comment"!*
• **Échantillon aléatoire simple** : prélèvement au hasard, de façon indépendante, d'un certain nombre d'éléments de la population statistique. Tous les éléments ont la même probabilité d'être inclus dans l'échantillon.

4.3. Caractère / Modalité / Variable:

4.3.1. Le caractère, les éléments d'un ensemble sont décrits par un caractère. Cela revient à établir une correspondance entre chaque élément i de l'ensemble E et l'ensemble X des modalités ou des valeurs du caractère. La fonction $f : E \rightarrow X \quad i \rightarrow x_i$ est une application au sens mathématique : chaque élément de E a une modalité (caractère qualitatif) ou une valeur (caractère quantitatif) et une seule dans X. Ainsi le caractère peut être défini comme une des caractéristiques ou des attributs d'un individu,

4.3.2. Modalité / Mesure : la modalité (respectivement la mesure) est l'une des formes particulière d'un caractère. Les différentes situations où les éléments de E *peuvent se trouver* à l'égard d'un caractère qualitatif considéré, sont les différentes **modalités** du caractère qualitatif X. Dans le cas ou le caractère X est quantitatif, les différentes situations où les éléments de E peuvent se trouver sont des **mesures**. Ces *modalités* ou ces *mesures* doivent être à la fois **incompatibles** (un élément de E ne peut prendre qu'une seule modalité) et **exhaustive** (à chaque élément de E doit pouvoir correspondre une modalité de X) de sorte que chaque élément de E ait une modalité et une seule dans X.

En statistique, chaque individu peut être définit par un ensemble d'attributs qui le caractérise dans le contexte d'un problème étudié.
La couleur du pelage est un caractère
Les variantes de la couleur du pelage sont des modalités : noir, gris,...
Le sexe est un caractère, ses modalités sont de 2 types : soit male soit femelle
La taille peut prendre plusieurs modalités : 1m ; 1,1m ; 1,2m etc…

Autres exemple de caractères :
Le taux de glycémie, la vitesse de coagulation ; la production laitière ;

4.3.3. Tableau élémentaire : c'est un tableau à simple entrée où les lignes correspondent aux éléments de l'ensemble étudié et les colonnes aux caractères (ou variables) décrivant ces éléments (Tableau 1 (1.1 et 1.2)).

Observations	Variables			
	Variable 1	Variable 2	Variable …	Variable p
Observation 1				
Observation 2				
Observation …				
Observation n				

***Tableau 1.1** : exemple de tableau de saisie élémentaire*
La première colonne est en principe réservée à la liste nominale des éléments.

a : tableau de mesures

	A	B	C
indiv. 1	3	110,5	55,22
indiv. 2	1	109,5	53
indiv. 3	2	108,7	57
indiv. 4	4	107,3	62,8
indiv. 5	0,5	102,1	61,2

b : tableau d'effectif (ou de %)

	A	B	C
indiv. 1	12	125	5
indiv. 2	11	130	6
indiv. 3	9	120	5
indiv. 4	13	110	4
indiv. 5	11	115	5

c : tableau de présence/absence

	A	B	C
indiv. 1	x	x	
indiv. 2	x		x
indiv. 3	x		x
indiv. 4		x	
indiv. 5		x	

d : tableau de contingence

	A	B	C
A	-		
B	1	-	
C	2	0	-

nombre d'objets possédant les caractères pris deux à deux

e : tableau de Burt

	B1	B2	B3
A1	3	2	0
A2	1	2	1
A3	2	0	4

tableau de contingence particulier, issu d'un tableau disjonctif complet et croisant les modalités des variables entre elles

f : tableau de similarité

	Sindiv.1	indiv. 2	indiv. 3
indiv. 1	-		
indiv. 2	1	-	
indiv. 3	0,6	0	-

calcul d'un indice de similarité entre individus

***Tableau 1.2** : Exemples de tableaux de données*

4.3.4. La variable statistique, chaque attribut (ou caractère ou caractéristique) a des modalités, ou peut s'exprimer selon une mesure, celles-ci varient d'un individu à l'autre ou d'un groupe d'individus à un autre groupe d'individus. La variable statistique est le nom que l'on donne à ces caractères (attributs, caractéristiques).

Explicitation de variable en biologie : caractéristique mesurable ou observable sur un élément (variable propre) ou dans son environnement (variable associée).

4.3.5. Nature des variables statistiques et échelles de mesures

Dans chaque étude statistique il est très important de considérer la nature des données (observations, caractères, attributs) que l'on va tester. D'elle dépend la nature des opérations possibles et donc des statistiques utilisables dans chaque situation. Il est donc primordial de préciser la nature de chaque variable, ou caractère. Il existe **deux types de variables** (ou observations, celles-ci peuvent être soit **quantitatives** soit **qualitatives**. Ces variables peuvent être mesurées d'où l'importance du **choix des échelles de mesures**, c'est-à-dire, des règles permettant d'affecter une valeur à chaque individu de la population ou de l'échantillon.

4.3.5.1. Variable quantitatif : c'est un caractère auquel on peut associer un nombre c'est-à-dire, pour simplifier, que l'on peut "mesurer" (grandeur mesurable). Les différentes situations où peuvent se trouver les éléments sont des *mesures*; elles sont ordonnables et la moyenne a une signification On distingue alors deux types de caractère quantitatif :

a - Variable discrète ou discontinue : c'est un caractère quantitatif, un tel caractère ne prend qu'un nombre fini de valeurs (valeur entière dénombrable et sans aucune valeur intermédiaire). Les différentes situations où peuvent se trouver les éléments (observations, mesures, valeurs,…) sont des nombres isolés dont la liste peut être établie a priori. Exemple: (nombre d'enfants, nombre de pétales d'une fleur, nombre de dents,..) : (1 ; 2 ; 3 ; 4 ; 5 ;….10 ; 11 ;…)

b - Variable continue : c'est un caractère quantitatif, un tel caractère peut, théoriquement, prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Toutes les valeurs ne sont pas dénombrables et ne peuvent pas être établit a priori. Ses valeurs sont alors regroupées **en classes** (taille, temps, poids, vitesse, glycémie, altitude, surfaces,….) (1,60 m ; 1,61 m ; 1,62 m ;…..)

c - Les mesures des données ou variables quantitatives comprennent les dénombrements (ou comptages) et les mesures (ou mensurations).

c_1 - Dans le cas des **dénombrements**, la caractéristique étudiée est une variable discrète ou discontinue, ne pouvant prendre que des valeurs entières non négatives (nombre de fruits par rameau, nombre de pétales par fleur, nombre de têtes de bétail..).

Il suffit de compter le nombre d'individus affectés par chacune des valeurs de la variable. Exemple : nombre de pétales par fleur dans un échantillon de 1000 fleurs de *Renonculus repens*.

Nombre de pétales par fleur	3	4	5	6	7
Nombre de fleurs	1	20	959	18	2

c_2 - Dans le cas des **mesures**, la variable est de nature continue (hauteur, poids, surface, concentration, température..). Les valeurs possibles sont illimitées mais du fait des méthodes de mesure et du degré de précision de l'appareil de mesure, les données varient toujours de façon discontinue.

Les mensurations peuvent être réalisées dans plusieurs échelles de mesure : l'échelle numérique, l'échelle de rapport, l'échelle d'intervalle. Elles sont manipulables suivant les opérations de l'arithmétique.

$c_{2.1}$ - L'**échelle numérique** est caractérisée par l'importance des valeurs mesurées. Le (0) signifie bien l'absence du phénomène. Exemple : population, taux de fécondité, précipitations.

$c_{2.2}$ - L'**échelle de rapport ou de taux** exprime le rapport entre deux valeurs. Leur total n'a pas de signification et caractérisé par l'existence d'un zéro absolu et de distances de taille connue entre deux valeurs quelconque de l'échelle. C'est le cas de la mesure de la masse ou du poids. En effet, les échelles de mesure des poids en pounds ou en grammes ont toutes deux un zéro absolu et le rapport entre deux poids quelconque d'une échelle est indépendant de l'unité de mesure (le rapport des poids de deux objets mesurés en pounds et celui de ces mêmes objets mesurés en grammes sont identiques). (Densité de population, proportion à une date ou à un lieu donnée).

$c_{2.3}$ - Dans l'**échelle d'intervalle**, le point zéro et l'unité de mesure sont arbitraires mais les distances entre deux valeurs quelconques de l'échelle sont de taille connue. Une telle échelle permet de repérer la position de chaque élément par rapport à une origine arbitraire. La valeur 0 est donc conventionnelle et ne signifie pas l'absence du phénomène C'est le cas de la mesure de la température (échelle Fahrenheit ou Celsius), de la Latitude de la Longitude, l'altitude, ...

Ces échelles quantitatives sont compatibles avec l'utilisation de tests paramétriques.

4.3.5.2. Variable qualitative : c'est un caractère qualitatif, dans ce type de variable les modalités ne sont pas quantifiables (pas mesurables) (couleur des yeux, douleur, …). Ce sont des noms ou ce qui revient au même des sigles ou des codes. Les différentes modalités ne sont pas ordonnables. Attention, même si les modalités sont des codes numériques, les opérations sur les modalités n'ont aucun sens.
Exemple : type de relief avec trois modalités (plaine, montagne, plateau), ou encore taille d'une niche écologique avec quatre modalités (petite, moyenne, grande, très grande). Les données qualitatives peuvent être assimilées au cas des variables discontinues, en supposant que les différentes variantes du caractère qualitatif sont rangées dans un ordre correspondant par exemple à la suite des nombres entiers positifs (différentes couleurs, différents degrés d'infection...). Les données qualitatives peuvent être réalisées dans deux échelles de mesure : échelle de rangement et l'échelle nominale. Ces données ne sont pas manipulables par l'arithmétique.

a - Dans l'échelle ordinale (de rangement), on parle dans ce cas de **caractère ordinal (caractères qui peuvent être** exprimés sur une échelle ordinale) : dans cette échelle chaque *modalité* est explicitement significative du rang pris par chaque individu pour le caractère considéré. Si E possède N éléments, les modalités seront 1^{er}, 2^{eme}, 3^{eme}, ... n^{eme}. Comme on possède juste l'ordre des individus, on ne sait rien de l'intervalle des valeurs. Il existe une certaine relation entre les objets du type plus grand que, supérieur à, plus difficile que, préférée à..... Une transformation ne changeant pas l'ordre des objets est admissible. La statistique la plus appropriée pour décrire la tendance centrale des données est la médiane.

b - Dans l'échelle nominale, les nombres ou symboles identifient les groupes auxquels divers objets appartiennent. C'est le cas des numéros d'immatriculation des voitures ou de sécurité

sociale (chaînes de caractères). Le même nombre peut être donné aux différentes personnes habitant le même département ou de même sexe constituant des sous-classes. Les symboles désignant les différentes sous-classes dans l'échelle nominale peuvent être modifiés sans altérer l'information essentielle de l'échelle. Les seules statistiques descriptives utilisables dans ce cas sont le mode, la fréquence... et les tests applicables seront centrés sur les fréquences des diverses catégories.

Ces deux dernières échelles ne permettent que l'utilisation de tests non paramétriques.

4.3.5.3. Exemple d'illustration des principaux types de descripteurs

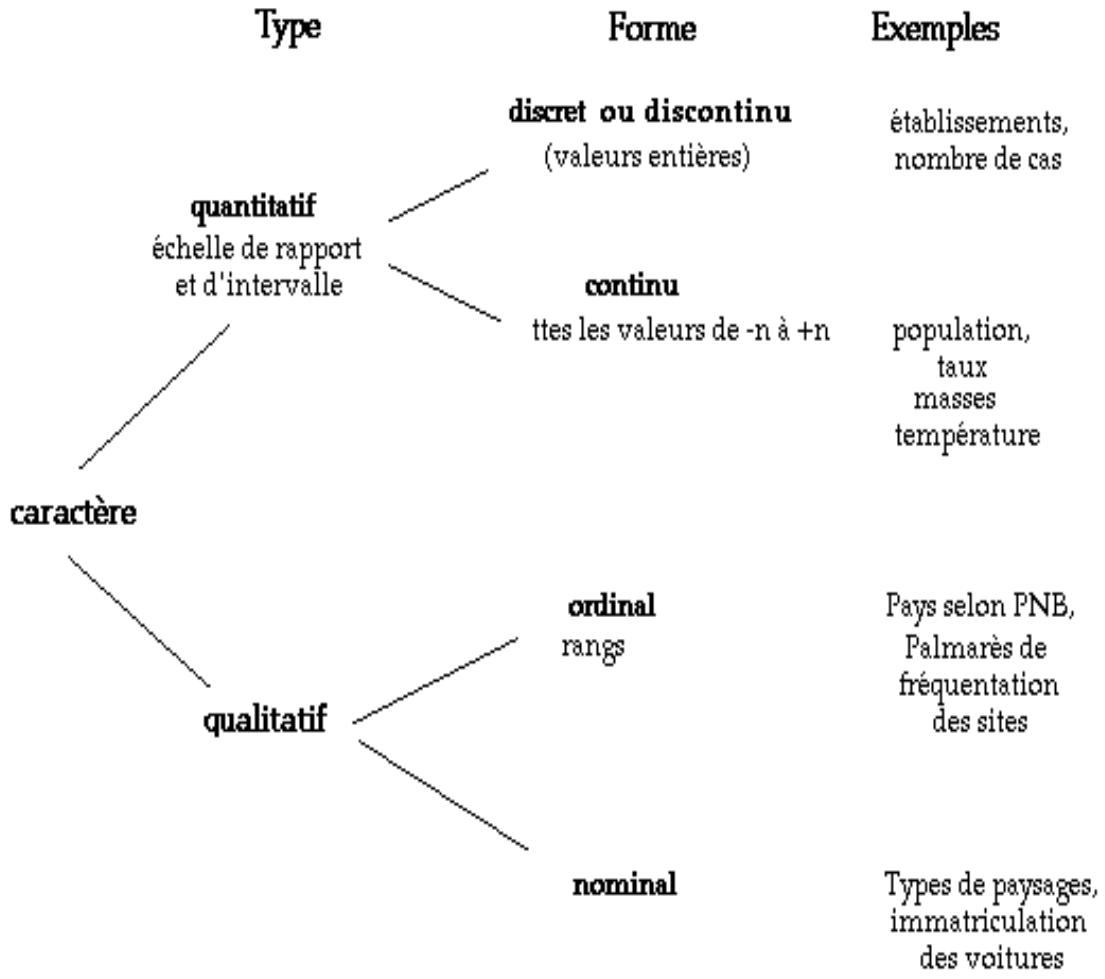


Figure 2 : *Typologie des caractères pour une approche statistique*

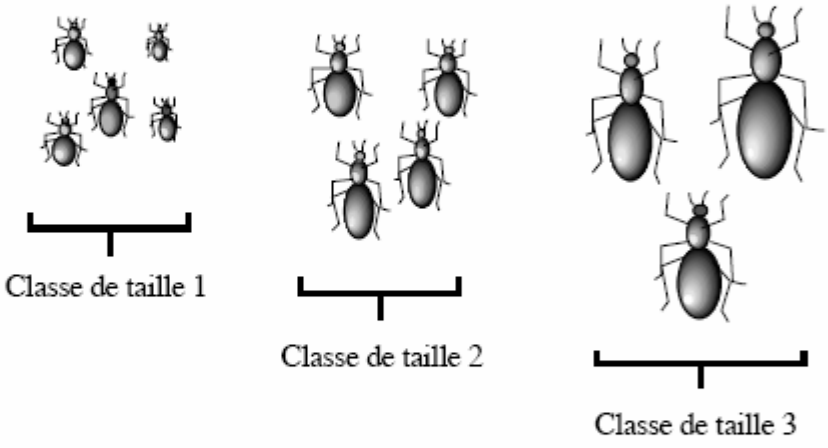
Binaires: 1 - 0 présent - absent

Descripteur "espèce" : Description "espèce présente": 1 ; Description "espèce absente": 0

	Relevé 1	Relevé 2	Relevé 3
Esp. 1	1	0	1
Esp. 2	1	1	0
Esp. 3	0	0	1

Multiples: - **non-ordonnés**, nominaux : ex. couleurs, type de sol...

- **ordonnés**: - **semi-quantitatifs**, ordinaux, de rang, : ex. classes de taille (0-10 cm, 10-50 cm, plus de 50 cm...), rang dans une course.



- **quantitatifs** : - **discontinus** (ex.: nombre de personnes dans cette salle, nb. d'individus par espèce...

	Relevé 1	Relevé 2	Relevé 3
Esp. 1	12	0	18
Esp. 2	3	56	0
Esp. 3	0	0	1

- **continus** (ex.: température, longueur, ...)

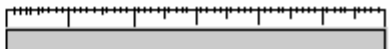


Figure 3 : *Exemples de variables statistiques et échelles de mesures*

APPLICATION I

Application I.1

Identifiez le type (et le sous-type) des variables suivantes :	**Réponses**
a) Le nombre d'animaux par laboratoire ;	a) quantitatif discret
b) La niche écologique principale ;	b) qualitatif nominal
c) Le modèle de matériel utilisé ;	c) qualitatif nominal
d) La distance en kilomètre entre le prélèvement A et le prélèvement B ;	d) quantitatif continue
e) Être végétarien ou non ;	e) qualitatif ordinal
f) Le temps passé à observer le comportement X ;	f) quantitatif continue
g) Avoir ou non une réponse;	g) qualitatif ordinal
h) Le nombre de frères et soeurs.	h) quantitatif discret

4.3.6. Variables dépendantes et indépendantes

En statistique on adopte encore une autre dichotomie pour le concept de variable en parlant de variables indépendantes et de variables dépendantes.

4.3.6.1. Les variables indépendantes sont celles qui sont manipulées par l'expérimentateur (l'appartenance au groupe et nous contrôlons les traitements appliqués aux différents groupes).

4.3.6.2. Les variables dépendantes sont celles qui sont mesurés, référencés, exemple de données (survie, résistances, tolérance, performance, …).

Fondamentalement, une étude porte sur les variables indépendantes et les résultats de l'étude (les données) sont les variables dépendantes.

4.3.7. La variabilité et l'incertain en biologie

Toutes les questions, proprement biologique en relation avec les statistiques, reflètent une propriété fondamentale des systèmes biologiques qui est leur variabilité. Cette variabilité est la somme d'une variabilité expérimentale (liée au protocole de mesure) et d'une variabilité proprement biologique. On peut ainsi décomposer la variabilité d'une grandeur mesurée en deux grandes composantes :

Variabilité Totale = Variabilité Biologique + Variabilité Métrologique

4.3.7.1 La variabilité biologique

Elle peut être décomposée en deux termes :

- d'une part la **variabilité intra-individuelle**, qui fait que la même grandeur mesurée chez un sujet donné peut être soumise à des variations aléatoires ;

- d'autre part la **variabilité interindividuelle** qui fait que cette même grandeur varie d'un individu à l'autre.

Variabilité Biologique = Variabilité intra-individuelle + Variabilité interindividuelle

La variabilité intra-individuelle peut être observée lors de la mesure de la performance d'un athlète qui n'est pas capable des mêmes performances à chaque essai, mais qui se différencie des autres athlètes (variabilité interindividuelle). En général, la variabilité intra est moindre que la variabilité inter.

4.3.7.2 La variabilité métrologique

Elle peut être elle aussi décomposée en deux termes : d'une part les conditions expérimentales dont les variations entraînent un facteur d'aléas ; et d'autre part les erreurs induites par l'appareil de mesure utilisé.

Variabilité Métrologique = Variabilité Expérimentale + Variabilité instrumentale (appareil de mesure)

La mesure de la pression artérielle peut grandement varier sur un individu donné suivant les conditions de cette mesure ; il est ainsi recommandé de la mesurer après un repos d'au moins 15 minutes, allongé, en mettant le patient dans des conditions de calme maximal. Cette recommandation vise à minimiser la variabilité due aux conditions expérimentales. La précision de l'appareil de mesure est une donnée intrinsèque de l'appareil, et est fournie par le constructeur.

4.3.8. Propriétés des variables

• Caractéristiques **mesurables** ou **observables.**
• **Propres** (attribut de l'élément) ou **associées** (composante de son environnement).
• **Aléatoires** (différentes variantes peuvent apparaître, chacune avec une certaine probabilité) ou
contrôlées (le chercheur obtient avec certitude la variante désirée, en général par manipulation).
• **Dépendantes** (on cherche à en comprendre ou prévoir le comportement) ou **indépendantes** (expliquent par hypothèse au moins une partie du phénomène étudié).
• **Simples** ou **complexes** (ex.: rapports, pourcentages...).

• Divers types mathématiques et échelles de variation.

4.4. Inférence et risque statistique

• **Inférence statistique**: généralisation à la population statistique des résultats d'un test statistique réalisé sur un échantillon représentatif de cette population. Cette généralisation se fait au risque du statisticien.

• **Généralisation à la population-cible**: lorsque cette dernière est différente de la population statistique, cette généralisation se fait au risque du biologiste.

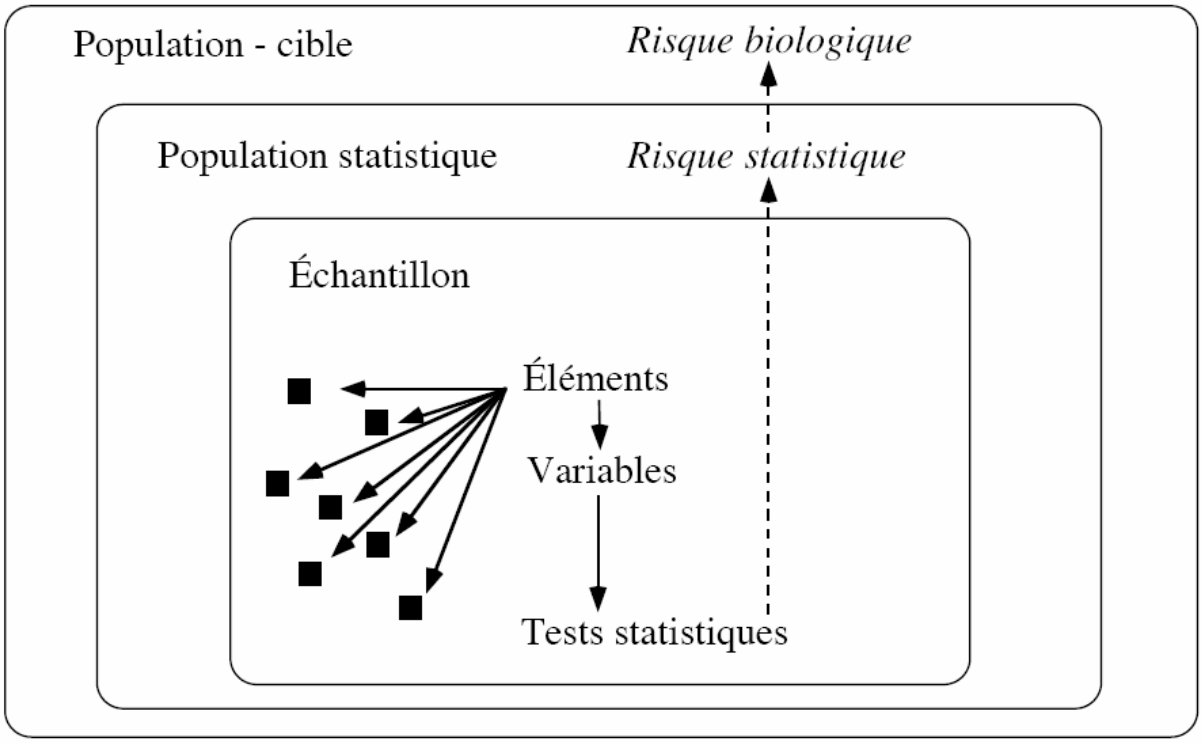


***Figure 4** : Représentation graphique de différentes populations et des risques*

4.5. Une définition plus explicite de la biostatistique :

La statistique est un ensemble d'instruments scientifiques par lesquels on recherche à expliquer certains phénomènes. Elle se compose de méthodes permettant de recueillir, de classer, de présenter et d'analyser des observations relatives à ces phénomènes pour en tirer ensuite des conclusions et prendre des décisions.

En biologie, la statistique est un ensemble de méthodes visant à décrire, à résumer et à interpréter des phénomènes dont le caractère essentiel est la variabilité.

4.6. Dénomination mathématique :

- E : représente l'ensemble de tous les individus sur lequel porte l'étude statistique

- Ω : représente la population sur laquelle porte l'étude statistique,
Si E est une énumération exhaustive de tous les individus susceptibles d'être analysés, il peut être appelé population ou univers et sera noté **Ω** .
Dans le cas contraire, E sera un échantillon de **Ω** .

- **|E|** : représente le cardinal de E, c'est le nombre de données (ou d'observations) référencées, il correspond à **l'effectif** ou **la taille** qui est généralement noté **N**. La taille de l'échantillon est l'effectif ou le nombre d'individus sur lequel sont réalisé effectivement les observations, c'est un sous ensemble de E (dans le cas où E caractérise la population entière), il correspond généralement au cardinal **|E|**.

- N : représente la taille d'une population ou d'un échantillon, c'est un nombre qui désigne le nombre d'individu que rassemble un échantillon ou une population.

- **p :** représente l'ensemble des variables caractérisant les individus sur lequel porte l'étude statistique.

E est de dimension p, si l'analyse de E est faite selon p variables (où p ∈ *IN(entier naturel)).*

Exemple : Dans une population E, on étudie 4 variables (w, x, y, et z) :
w : age, **x :** le sexe, **y :** la taille et **z :** le poids. Dans ce cas E est de dimension p, ou de dimension 4

- Système de notation
• Lorsqu'on mesure la valeur ou observe l'état d'un certain nombre de variables sur un élément, on utilise l'une ou l'autre des notations suivantes pour désigner les variables :
- s'il y a une, deux ou trois variables : *x, y* et *z*
- s'il y a plus de trois variables: *x1, x2, x3 ... xj ... xp*
Les variables sont donc numérotées de la première, à la *p*-ième, une variable quelconque étant la *j*ième.
• Un jeu de données (p.ex. un échantillon) comporte n éléments. Un quelconque de ces éléments est le *i*-ième. Ces éléments sont souvent qualifiés d'**observations** ou d'**objets**.
- lorsqu'on mesure la valeur d'une variable *x* sur un élément quelconque (le *i*-ième), on désigne cette valeur par *xi.*
- *i* varie de 1 à *n*, donc on a les mesures *x1, x2, ... xi ... xn*.
• Si le jeu de données consiste en un tableau de *n* éléments décrits par *p* variables (tableau *n* × *p*), on note:

	Variables						
Observations	x_1	x_2	x_3	…	x_j	…	x_p
1	x_{11}	x_{12}	x_{13}	…	x_{1j}	…	x_{1p}
2	x_{21}	x_{22}	x_{23}	…	x_{2j}	…	x_{2p}
3	x_{31}	x_{32}	x_{33}		x_{3j}		x_{3p}
…							
i	x_{i1}	x_{i2}	x_{i3}	…	x_{ij}	…	x_{ip}
…							
n	x_{n1}	x_{n2}	x_{n3}	…	x_{nj}	…	x_{np}

• Il arrive que les éléments soient répartis en *k* groupes caractérisés par une variable qualitative. Dans ce cas, on peut aussi noter les observations d'une variable par un double indice, le premier désignant le numéro de l'observation au sein d'un groupe (*i*-ième élément), le deuxième désignant le numéro du groupe (*g*ième groupe ou *j*-ième groupe):
- x_{ig} ou encore x_{ij} la mesure prise sur le ***i*-ième** élément du ***g*-ième** (ou ***j*-ième**) groupe.

- Notation somme $\sum$ (sigma)

• La lettre grecque sigma majuscule $\sum$ désigne une sommation (addition de tous les éléments d'un ensemble).

• La sommation des valeurs des *n* observations d'une variable *x*, soit de tous les *xi* pour *i* allant de 1 à *n*, se note: $\sum_{i=1}^{n} x_i$

Donc $\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_i .. + x_n$

Si l'on veut faire la somme de tous les éléments d'un tableau de *n* observations et *p* variables, on écrira:

$$\sum_{i=1}^{n}\sum_{j=1}^{p} xij = x_{11} + x_{12} + ... + x_{1i} .. + x_{n1} + x_{12} + x_{22} + ... + x_{ij} .. + x_{np}$$

• Si ***a*** est une constante, tous les a_i sont égaux. Donc:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + ... + a_i .. + a_n = na$$

$$\sum_{i=1}^{n} (x_i + a) = na + \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

On se sert de ces propriétés des sommations pour **simplifier** ou **développer** des expressions comportant des sommations.

Application I.2

• **Simplifier** le plus possible l'expression suivante:

$$ax_1 + ax_2 + ... + ax_n - ay_1 - ay_2 - ... - ay_n$$

On peut écrire:

$$\sum_{i=1}^{n} ax_i - \sum_{i=1}^{n} ay_i = \sum_{i=1}^{n} (ax_i - ay_i) = \sum_{i=1}^{n} a(x_i - y_i) = a\sum_{i=1}^{n} (x_i - y_i)$$

Ou encore:

$$\sum_{i=1}^{n} ax_i - \sum_{i=1}^{n} ay_i = a\sum_{i=1}^{n} x_i - a\sum_{i=1}^{n} y_i = a\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right) = a\sum_{i=1}^{n} (x_i - y_i)$$

• **Développer** le plus possible l'expression suivante:

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

On peut écrire:

$$\sum_{i=1}^{n} (2a + 2x_i) = 2a + 2x_1 + 2a + 2x_2 + ... + 2a + 2x_n$$

Ou encore:

$$2\sum_{i=1}^{n} (a + x_i) = 2\left(na + \sum_{i=1}^{n} x_i\right) = 2na + 2\sum_{i=1}^{n} x_i = \underbrace{2a + 2a + 2a + ...}_{n \text{ termes}} 2x_1 + 2x_2 + ... + 2x_n$$

- Fréquences absolues, relatives et cumulées (voir tableau exemple)

Désigné par « **F** » ou « **f** » La notion de fréquence peut être exprimée de plusieurs manières :
* Fréquence absolue (effectif)
* Fréquence relative (ou fréquence)
* Fréquences cumulées

	Variables				
Exemples de Fréquences	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	**Total**
Effectif ou Fréquence absolue (n_i)	**8**	**2**	**9**	**3**	**22**
Fréquence absolue cumulée croissante	**8**	8+2=**10**	10+9=**19**	19+3=**22**	
Fréquence absolue cumulée décroissante	**22**	22-8=**14**	14-2=**12**	12-9=**3**	
Fréquence relative (f_i)	**8/22**	**2/22**	**9/22**	**3/22**	22/22 = **1**
Fréquence relative cumulée croissante	**8/22**	8/22+2/22=**10/22**	**19/22**	**22/22**	
Fréquence relative cumulée décroissante ou fréquence cumulée décroissante	22/22 = **1**	22/22-8/22 =**14/22**	(14-2)/22 **=12/20**	(12-9)/22 **=3/22**	

Tableau 2 : *Exemples explicatifs des fréquences*
(Ce tableau servira d'exemple pour comprendre les notions de fréquences)

*** Fréquences absolues = Effectif**

Le terme de **fréquence absolue** désigne les effectifs : a chaque modalité x_i du caractère X, peut correspondre un ou plusieurs individus dans l'échantillon de taille n. On appelle **effectif (ou fréquence absolue)** de la modalité x_i**,** le nombre n_i où n_i est le nombre d'individu de chacune des modalités

*** Fréquence relative = Fréquences**

On appelle **fréquence** de la modalité x_i**,** le nombre f_i tel que

$$f_i = \frac{n_i}{n}$$

Remarques :

Rq$_1$: Le pourcentage est une fréquence exprimée en pour cent. Il est égal à 100 f_i.

Rq$_2$: L'emploi des fréquences ou fréquences relatives s'avère utile **pour comparer deux distributions** de fréquences établies à partir d'échantillons de **taille différente**.

*** Fréquences cumulées = fréquences relatives cumulées**

On appelle **fréquences cumulées** ou **fréquences relatives cumulées** en x_i, le nombre

$f_i$cum tel que $f_i cum = \sum_{p=1}^{i} f_p$

Remarques

Rq1 : la taille de l'échantillon est = $\sum_{i=1}^{n} n_i = n$

Rq2 : $\sum_{i=1}^{k} f = 1$

- Fonctions cumulatives
* fonction cumulée croissante ou ascendante
* fonction cumulée décroissante ou descendante

Définitions
Soit *S* une série statistique à une variable de type quantitatif et *a* une modalité de *S*. La **fréquence cumulée croissante** associée à *a* est la somme des fréquences de toutes les modalités inférieures ou égales à *a* dans la série *S*.
Dans le cas d'une série *S* dont les modalités sont regroupées en classes, la **fréquence cumulée croissante** de la classe [*a* ; *b*[est la somme des fréquences de cette classe et des classes qui précèdent (c'est-à-dire dont les éléments sont strictement inférieurs à *a*) s'il y en a.

Remarques
* La fréquence cumulée croissante de la plus petite modalité ou de la classe à laquelle appartiennent les plus petites modalités est égale à la fréquence de cette modalité ou de cette classe;
* La fréquence cumulée croissante de la plus grande modalité ou de la classe à laquelle appartiennent les plus grandes modalités est égale à 1 (ou à 100 % pour les fréquences exprimées en pourcentages).

5. REPRESENTATION DES DONNEES

Il existe plusieurs niveaux de description statistique : la présentation brute des données, des présentations par tableaux numériques, des représentations graphiques et des résumés numériques fournis par un petit nombre de paramètres caractéristiques.

Nous reviendrons sur les représentations graphiques et les tableaux respectivement dans les paragraphes suivants et dans les exemples

5.1. Tableaux statistiques
En général une série statistique à caractère discret se présente sous la forme :

Valeurs	X1	X2	…..	Xp
Effectifs	N1	N2	…..	Np
Fréquences	F1	F2	…..	Fp

Plutôt que réécrire ce tableau on écrira souvent : la série (xi, ni). (On n'indique pas le nombre de valeurs lorsqu'il n'y a pas d'ambigüité). Souvent on notera N l'effectif total de cette série donc $N = n1+n2+. \ . \ .+np$. (*Voir paragraphe 4.3.3 ; 4.6 Tableau 1 et 2)*

Application I.3 : Caractères quantitatifs discrets

Dans le cas d'un **caractère quantitatif discret,** l'établissement de la distribution des données observées associées avec leurs fréquences est immédiat.

Exemple :

La **cécidomyie** du hêtre provoque sur les feuilles de cet arbre des galles dont ***la distribution de fréquences observées*** est la suivante :

Caractère *X* : ***x_i :*** nombre de galles par feuille	0	1	2	3	4	5	6	7	8	9	10
n_i : nombre de feuilles portant x_i galles	182	98	46	28	12	5	2	1	0	1	0
f_i : fréq. relative	0,485	0,261	0,123	0,075	0,032	0,013	0,005	0,003	0	0,003	0
f_i cum. : fréq. Relative cumulée	0,485	0,746	0,869	0,944	0,976	0,989	0,994	0,997	0,997	1	1

La taille de l'échantillon étudié est n = 375 feuilles

Application I.4 : Utiliser le logiciel Excel pour dresser ces tableaux et réaliser les calculs

Application I.5 : Caractères quantitatifs continus

(**Mots clés** : Nombre de classes, intervalle entre classe (amplitude), étendu de la variable X)

Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable **une répartition en classes** des données. Cela nécessite de définir **le nombre de classes** attendu et donc **l'amplitude** associée à chaque classe ou **intervalle de classe**.

En règle générale, on choisit des classes de même amplitude. Pour que la distribution en fréquence est un sens, il faut que chaque classe comprenne un nombre suffisant de valeurs (n_i).

Diverses formules empiriques permettent d'établir le **nombre de classes** pour un échantillon de taille n.

La règle de **STURGE** : $$Nombre\,de\,classe = 1 + (3{,}3 \log n)$$

La règle de **YULE :** $$Nombre\,de\,classe = 2{,}5 \sqrt[4]{n}$$

L'**intervalle** entre chaque classe est obtenu ensuite de la manière suivante :

Intervalle de classe = (*X* max - *X* min) / Nombre de classes

Avec $\boldsymbol{X}_{max}$ et $\boldsymbol{X}_{min}$, respectivement la **plus grande** et la **plus petite valeur de *X*** dans la série statistique.

A partir de $\boldsymbol{X}_{min}$ on obtient **les limites de classes** ou **bornes de classes** par addition successive de l'intervalle de classe. En règle général, on tente de faire coïncider **l'indice de classe** ou valeur centrale de la classe avec un nombre entier ou ayant peu de décimales. Toutes les données sont comprises entre X_{min} et X_{max} et chaque donnée appartient à une et une seule classe.

Exemple :

Dans le cadre de l'étude de la population de **gélinottes huppées** (Bonasa *umbellus*), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

158	152	171	163	140	157	162	171	158	164	163	159	153
160	149	158	152	165	156	162	150	154	155	162	155	164
164	157	159	158	159	153	163	158	174	162	156	151	
160	158	162	166	162	164	158	153	165	158	150	160	

• **Définition du nombre de classes :**

Règle de Sturge : 1 + (3,3 log 50) = 6,60

Règle de Yule : $2{,}5\sqrt[4]{50}$ = 6,64

Les deux valeurs sont très peu différentes

• **Définition de l'intervalle de classe :**

$IC = \frac{174-140}{6{,}6} = 5{,}15mm$ que l'on arrondit à 5 mm par commodité

• **Tableau de distribution des fréquences**

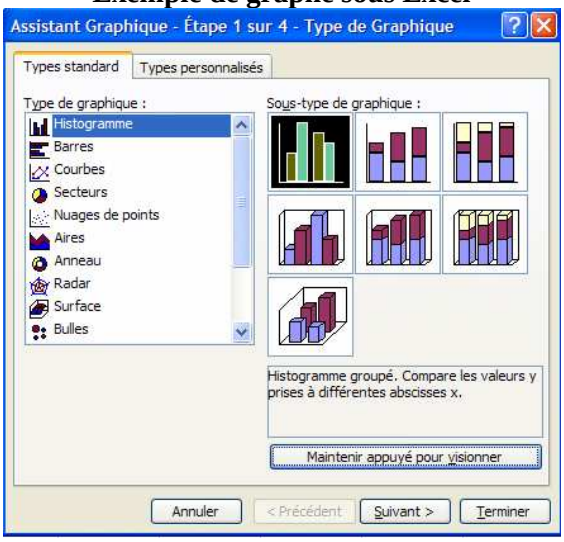
Caractère *X* : x_i : longueur de la rectrice bornes des classes en mm	**[140-145[**	**[145-150[**	**[150-155[**	**[155-160[**	**[160-165[**	**[165-170[**	**[170-175[**
Valeur médiane des classes x_i'	**142,5**	**147,5**	**152,5**	**157,5**	**162,5**	**167 ,5**	**172,5**
n_i **:** nombre d'individu par classe de taille x_i	**1**	**1**	**9**	**17**	**16**	**3**	**3**
f_i : fréquence relative	**0,02**	**0,02**	**0,18**	**0,34**	**0,32**	**0,06**	**0,06**
f_i ***cum.*** : fréquence relative cumulée (croissante)	**0,02**	**0,04**	**0,22**	**0,56**	**0,88**	**0,94**	**1**

Application I.6 : Utiliser le logiciel Excel pour dresser ce tableau, calculer l'intervalle des classes et réaliser les calculs

5.2. Représentations graphiques et statistique descriptive

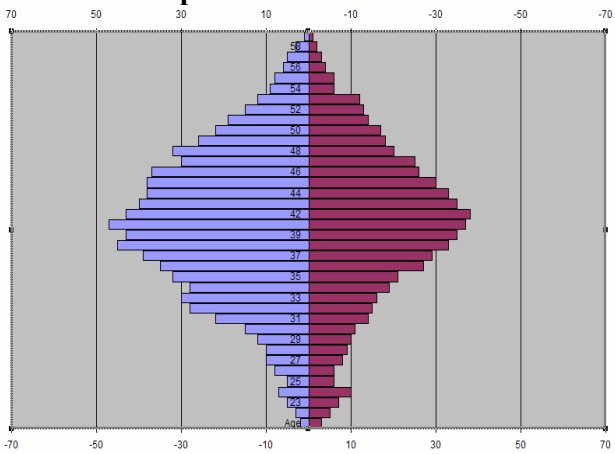
Les représentations graphiques sont très importantes en statistique descriptive. Elles ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies. La représentation graphique des données montre la forme générale de la distribution et donne une image de la grandeur des nombres qui constituent les données. D'autres statistiques simples sont utilisées pour représenter le centre de la distribution et les mesures liées à la dispersion des observations autour de cette tendance centrale.

Exemple de graphe sous Excel



Dans cette partie, nous ne présenterons que les cas particuliers de l'histogramme et des **Barres à moustaches (Box Plot)** cependant, d'autres représentations seront abordées dans les différentes parties de ce fascicule. De plus, plusieurs activités pratiques de construction (voir TD) expliciteront les constructions de plusieurs types de graphes et présenterons leurs nombreux avantages (pour plus d'informations consulter le document « représentation graphique » du TD).

Exemple de construction en TD



5.2.1. L'histogramme

• Définition : L'histogramme consiste à faire figurer les effectifs d'une variable par classe de valeur.
• Il est représenté quand la variable est quantitative continue par des rectangles dont la surface (et non la hauteur) est proportionnelle aux effectifs.

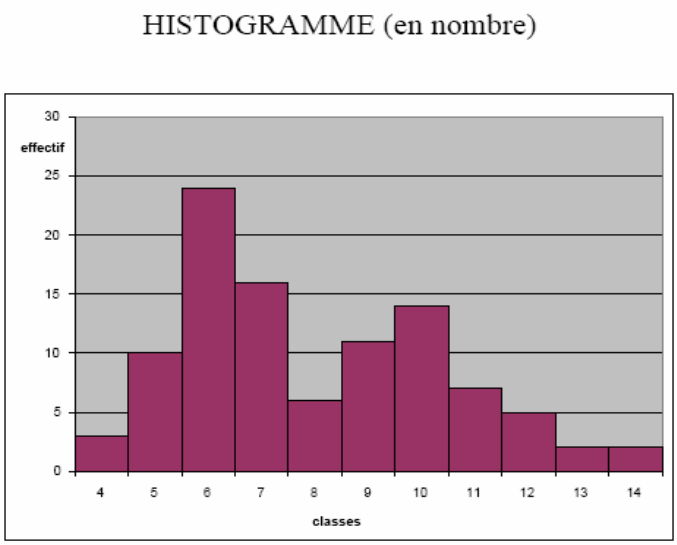
APPLICATION II

Application II.1 - L'histogramme : exemple

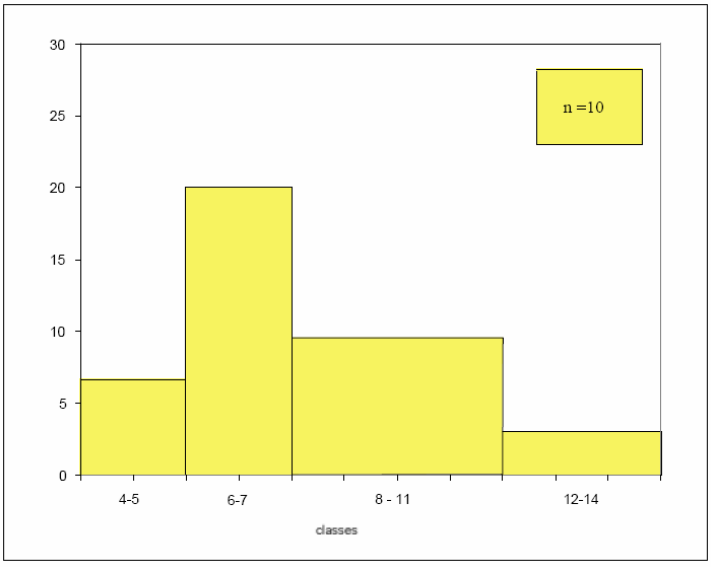
A partir de la liste des valeurs du tableau des effectifs suivante, construire l'histogramme correspondant (utiliser le logiciel Excel)

classes (mettre l'unité)	effectif (en nombre)
4	3
5	10
6	24
7	16
8	6
9	11
10	14
11	7
12	5
13	2

Tableau d'effectifs

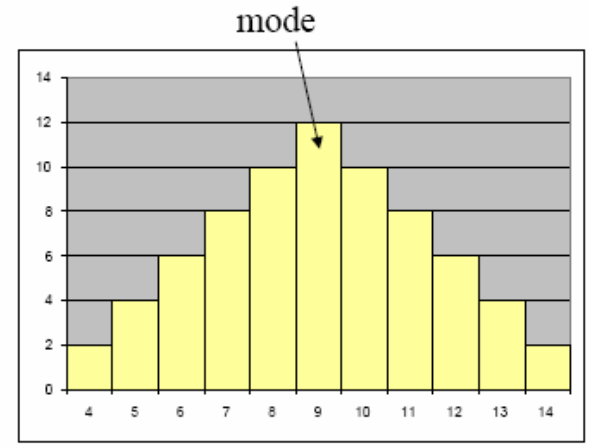


Application II.2 : Les classes peuvent être définies d'intervalles égaux ou non.
Dans ce dernier cas, seule la surface sera proportionnelle à l'effectif (et non la hauteur)

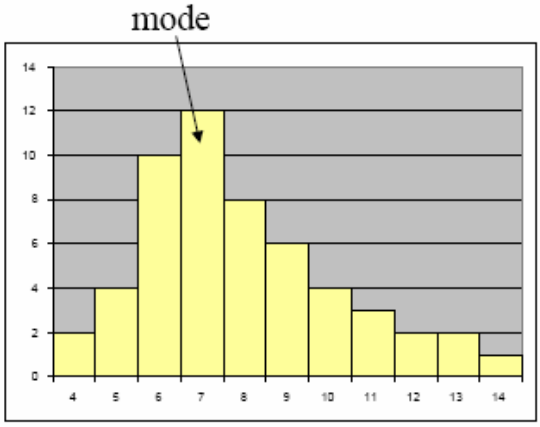


5.2.1.1. L'histogramme : paramètres de description

Pour décrire la forme d'un histogramme on utilise les notions de mode et de symétrie :
• le mode est la valeur dominante, dont l'effectif est le plus élevé. Un histogramme peut avoir aucun, un ou plusieurs modes. Dans un histogramme, le mode est le rectangle qui a l'aire la plus grande.

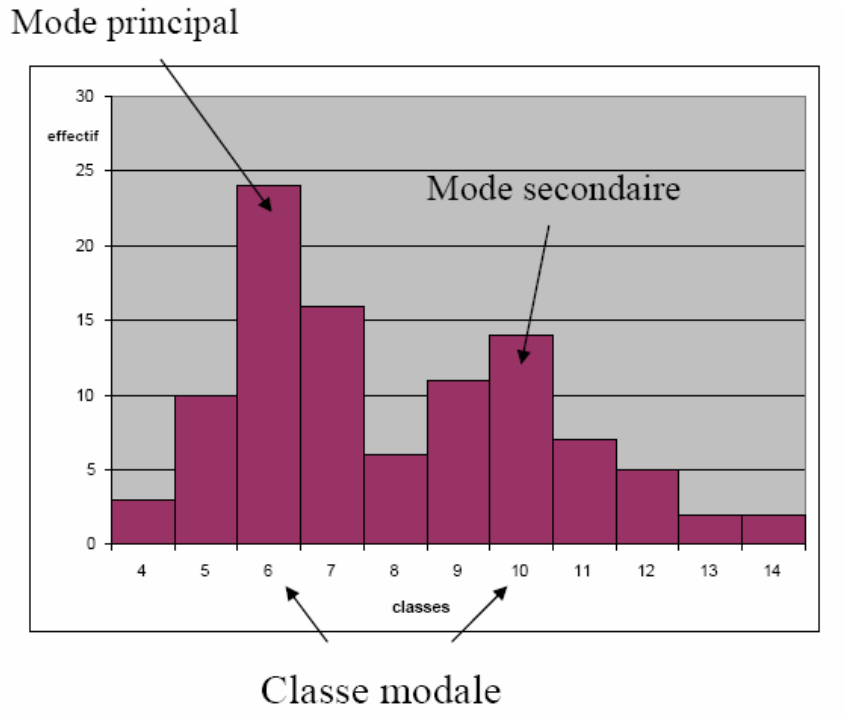


Distribution unimodale symétrique



Distribution unimodale asymétrique

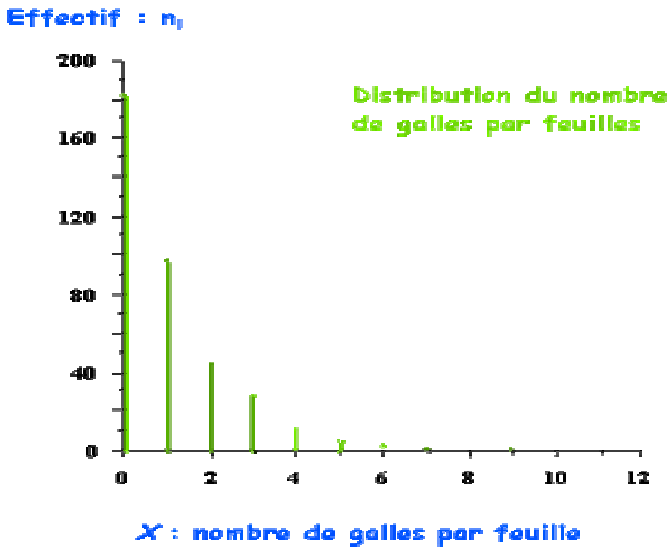
• La symétrie traduit la distribution des valeurs de part et d'autre du ou des modes
Cas d'une distribution bimodale asymétrique



<u>Application II.3 :</u> Caractères quantitatifs discrets

Pour les caractères quantitatifs discrets, la représentation graphique est le **diagramme en bâtons** où la hauteur des bâtons correspond à l'effectif n_i associé à chaque modalité du caractère x_i.

Exemple :

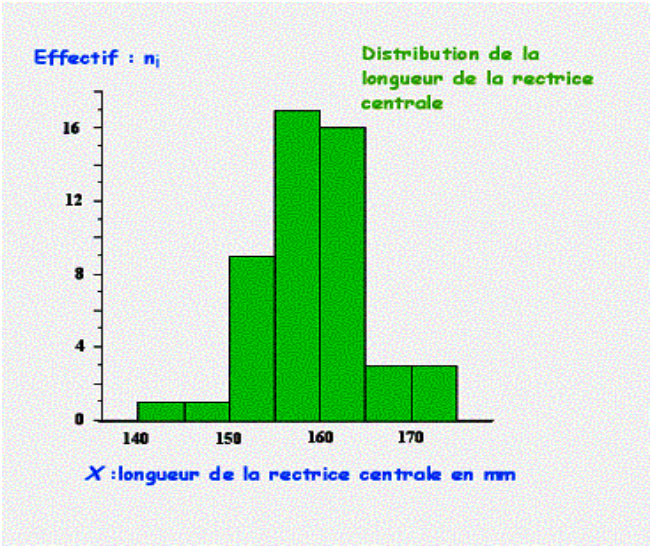


Dans l'exemple de la **cécidomyie** du hêtre, la distribution des fréquences observées du nombre de galles par feuille peut être représentée par **un diagramme en bâtons** avec en ordonnée les **effectifs** n_i et en abscisse les différentes **modalités** de la variable étudiée.

Application II.4 : Caractères quantitatifs continus

Pour les caractères quantitatifs continus, la représentation graphique est **l'histogramme** où la hauteur du rectangle est proportionnelle à l'effectif n_i. Ceci n'est vrai que si l'intervalle de classe est constant. Dans ce cas l'aire comprise sous l'histogramme s'avère proportionnelle à l'effectif total. En revanche lorsque les intervalles de classe sont inégaux, des modifications s'imposent pour conserver cette proportionnalité. Dans ce cas, en ordonnée, au lieu de porter l'effectif, on indique le rapport de la fréquence sur l'intervalle de classe. Ainsi la superficie de chaque rectangle représente alors l'effectif associé à chaque classe.

Exemple :



-

Dans l'exemple de la longueur de la rectrice centrale des individus mâles de la gélinotte huppée, la distribution des fréquences observées est représentée par un **histogramme** avec en ordonnée les **effectifs** n_i et en abscisse **les limites de classe** de la variable étudiée.

Application II.5 : Utiliser le logiciel Excel pour réaliser l'histogramme de l'application 2

5.2.2. Barre à moustache - Box Plot

Remarque : Pour comprendre cette partie il est nécessaire de se référer au paragraphe « 6.1. Paramètre de position et valeurs centrales ».

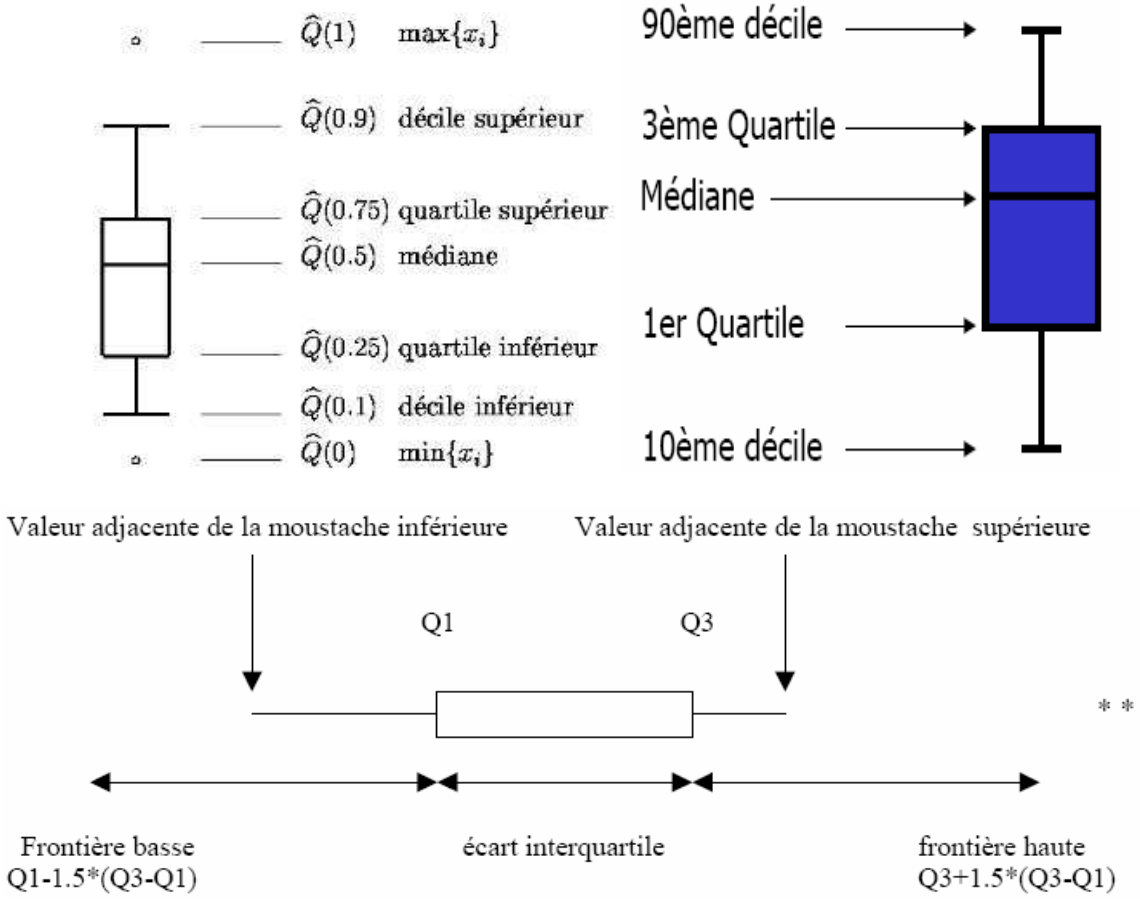
Une "boîte à moustaches" (traduction française du terme "Box and Whiskers Plot", ou en abrégé "Box Plot") est une représentation graphique de quelques paramètres de distribution d'une variable, inventée par Tukey en 1977. C'est une représentation graphique d'une variable quantitative qui permet d'appréhender (résumer une distribution empirique) la dispersion d'un échantillon.

Rappel

(En référence au paragraphe « 6.1. Paramètre de position et valeurs centrales ».)

On appelle intervalle inter-quantiles l'intervalle [Q (0,25), Q (0,75)], qui contient la moitié centrale des valeurs de l'échantillon. On appelle intervalle inter-déciles l'intervalle [Q (0,1), Q (0,9)], qui contient 80% des valeurs centrales de l'échantillon.

Ces intervalles sont à la base d'une représentation très compacte de la distribution empirique : le diagramme en boite (ou boîte à moustaches, box plot, box-and-whisker plot). Il n'y a pas de définition standardisée de cette représentation. Elle consiste en une boîte rectangulaire, dont les deux extrémités sont les quartiles. Ces extrémités se prolongent par des traits terminés par des segments orthogonaux (les moustaches). La longueur de ces segments varie selon les auteurs. Il existe donc plusieurs variantes pour représenter les boites à moustache, nous traiterons de celle la plus fréquemment utilisée. Nous proposons par exemple de fixer la longueur des segments aux déciles extrêmes. On représente aussi la médiane par un trait dans la boîte, et parfois les valeurs extrêmes par des points (voir figure ci-dessous).



Exemples de représentations graphiques: boîtes à moustache (box-plot)

Analysons rapidement les boites à moustaches ci dessus :
La boîte à moustaches utilise généralement 5 valeurs qui résument des données :
- Le minimum,
- Les 3 quartiles Q1, Q2 (médiane) et Q3,
- Le maximum.
Les quartiles Q1, Q2, Q3 sont les éléments essentiels de ce type de graphique. **Nous détaillerons les étapes de la construction des quartiles et de l'écart interquartile en TD.**

- **les premier et troisième quartiles** (Q1 (0,25) et Q3 (0,75)) : bordures inférieure et supérieure de la boîte rectangulaire
- **la médiane :** trait horizontal long au sein de la boîte rectangulaire (Q2 (0,5))
- **la moyenne :** marque plus (+) au sein de la boîte, pouvant être confondue avec la médiane
- **les extrémités inférieure et supérieure des moustaches** : marques en forme de tiret (-) située sur le trait vertical, et correspondant respectivement à la plus petite donnée supérieure à une valeur a_1, et à la plus grande donnée inférieure à une valeur a_3;

Il est possible de calculer ces extrémités avec les formules suivantes :

$\mathbf{a_1}$= Q1-1.5*(Q3-Q1) = Q1-1,5QI

avec QI = Intervalle inter-quartiles

$\mathbf{a_3}$= Q3+1.5*(Q3-Q1) = Q3-1,5QI

- **les minimum et maximum** : marques extrêmes en forme de cercle (o) ; si le minimum ou le maximum n'est pas confondu avec le tiret d'extrémité de moustache, c'est qu'il s'agit d'une valeur atypique ("outlier"). Les valeurs atypiques peuvent être situées strictement en dessous de la moustache inférieure $\mathbf{a_1}$ (nb atyp. inf.) ou strictement en dessus de la moustache supérieure $\mathbf{a_3}$ (nb atyp. sup.).

Pour plus de détails sur l'utilisation des boîtes à moustaches, voir TD.

Application II.6 : Création et test d'une macro BoxPlot sous Excel voir TD.

Remarque :
La médiane est une valeur centrale de l'échantillon : il y a autant de valeurs qui lui sont inférieures que supérieures. Si la distribution empirique de l'échantillon est peu dissymétrique, comme par exemple pour un échantillon simulé à partir d'une loi uniforme ou normale, la moyenne et la médiane sont proches. Si l'échantillon est dissymétrique, avec une distribution très étalée vers la droite, la médiane pourra être nettement plus petite que la moyenne. Contrairement à la moyenne, la médiane est insensible aux valeurs aberrantes. Elle possède une propriété d'optimalité par rapport à l'écart absolu moyen.

6. STATISTIQUES DESCRIPTIVES UNIVARIEES

Quelques exemples

*** Exemples de séries univariées**

Une série univariée est formée par une série de mesures d'une variable quantitative, généralement continue (valeurs décimales), effectuées sur un même échantillon :

Exemple 1 :

- 5 mesures du poids d'un organe (en g) : 14,5 13,2 18,63 15,0 13,33

Eventuellement : on peut avoir à faire un variable quantitative discrète (valeurs entières), pourvu que la notion de moyenne ait un sens par rapport à cette variable (ce n'est pas un "code"):

Exemple 2 :

- 7 mesures du "nombre de poils aux pattes d'une mouche" : 27 28 25 21 28 19 20

*** Données de deux séries univariées**

Dans cette situation, les deux séries de données concernent la même variable. Dans la situation la plus courante, la première série provient d'un échantillon "témoin", la seconde d'un échantillon "traité" :

Exemple 1 :

- On dispose de deux échantillons de rats males, dont on a mesuré le poids corporel (en g):
  - 3 rats TEMOINS : 410 432 417
  - 5 rats TRAITES par un anabolisant : 435 482 457 502 473

Autre situation fréquente : on observe le même échantillon "avant" et "après" un traitement :

Exemple 2 :

- On mesure l'hématocrite (unités arbitraires) avant et après un traitement anticoagulant :
  - les mesures AVANT le traitement : 97 103 95,5 102 100
  - les mesures APRES le traitement : 84 78 90,5 85 76

On peut aussi comparer des échantillons qui diffèrent par l'origine bio-géographique, l'âge, le sexe...

*** Données de plusieurs séries univariées**

Les différentes séries de données concernent la même variable. Plusieurs traitements ont été appliqués (ou bien on a échantillonné des populations réputées différentes) :

Exemple :

- On dispose de trois échantillons de rats males, dont on a mesuré le poids corporel (en g):
  - 3 rats TEMOINS : 410 ; 432 ; 417
  - 5 rats TRAITES par un anabolisant : 435 ; 482 ; 457 ; 502 ; 473
  - 4 rats traités par un PLACEBO : 422 ; 437 ; 395 ; 412

Les statistiques descriptives visent à représenter des données dont on veut connaître les principales caractéristiques quantifiant leur variabilité.

Trois aspects sont essentiels à l'interprétation d'une distribution :

- **Paramètre de position** : le centre de la distribution et la répartition autour d'une valeur centrale (moyenne, mode, médiane, quantiles, ..)
- **Paramètre de dispersion ou d'étendue** : les valeurs sont-elles dispersées ou concentrées ?
- **Paramètre de forme** : la forme de la distribution : la symétrie, l'aplatissement

6.1. Paramètre de position et valeurs centrales

Le but des valeurs centrales est de résumer en une seule valeur l'ensemble des valeurs d'une distribution statistique. Il existe quatre valeurs de positions :

- **Le mode** (Mo),
- **La moyenne** ($\overline{X}$ ou μ)
- **La médiane ou le médian** (Me ou Md)
- **Les fractiles** (Quantiles) (Qn)

Parmi ces valeurs les trois premières sont des valeurs de **position centrales** :

6.1.1. Le mode, ou valeur dominante, est la valeur la plus fréquente d'une distribution. Cette valeur se calcule toujours à partir d'un dénombrement des modalités du caractère. Il faut donc distinguer le cas des caractères discrets et des caractères continus (voir notions de bases).

*** Caractère qualitatif et caractère discret** : Pour un caractère qualitatif, ou pour un caractère quantitatif discret ayant un nombre de modalités inférieur au nombre d'éléments, le mode est la modalité ou la valeur qui a **la fréquence simple la plus élevée** (ou l'effectif le plus élevé, ce qui revient au même).

*** Caractère quantitatif continu :** Les modalités étant en nombre infini, il est peu probable que deux éléments aient la même valeur. Dans ce cas, le mode ne peut pas être défini directement, il faut au préalable établir une partition en classes. Le mode est alors le centre de la classe modale, c'est à dire de la classe qui a **la fréquence moyenne la plus élevée**.
Le mode correspond à la valeur lue en abscisse du sommet de l'histogramme. Lorsque celui-ci présente deux pics séparés par un creux, on dit que la distribution est **bimodale**.

APPLICATION III

Application III. 1 : Cas de calcul des modes :

- Cas 1 : Données rangées : le mode est la valeur de la donnée qui apparaît le plus fréquemment (celle qui a le plus d'occurrences) :
140 ; 141 ; 144 ; 144 ; 148 ; 148 ; 152 ; 152 ;152 ; 154 ; 155 ; 158 ; 158 ; 161 ; 170 ; 172
Le mode est 152 car il possède le plus grand nombre d'occurrences (il est référencé 3 fois)

- Cas 2 : Données condensées : le mode est la valeur de la donnée qui possède la fréquence la plus élevée (relative ou absolue).

Modalités xi (age en années)	**14**	**16**	**18**	**21**	**22**	**24**	**25**	**Total**
Fréquences absolues	5	**12**	10	8	11	7	3	56
Fréquences relatives	0,089	**0,214**	0,179	0,143	0,196	0,125	0,054	1,000

Dans cette série statistique, le mode est égal à Mo = 16 ans

- Cas 3 : Données groupées en classes : la classe modale est la classe ayant la plus haute fréquence (relative ou absolue).

Dans le tableau des classes relatives à la longueur de la rectrice de Bonasa *umbellus* , la classe modale est [155mm-160mm[. Il est possible de calculer de façon plus précise le mode en appliquant la formule suivante :

$$Mo = bmo + (\frac{\Delta 1}{\Delta 1 + \Delta 2})Lmo$$

bmo : Borne inférieur de la classe modale
Lmo : largeur de la classe modale

Δ1 = différence entre l'effectif de la classe modale et l'effectif de la classe précédente.

Δ2 = différence entre l'effectif de la classe modale et l'effectif de la classe qui suit.

Δ1 = (17-9) = 8 ; **Δ2** = (17-16) = 1 ; **bmo** = 155 ; **Lmo** = 5

$$Mo = 155 + (\frac{8}{8+1})5 = 159 \text{ mm}$$

6.1.2. La moyenne

Formalisation mathématique de la moyenne arithmétique

La moyenne arithmétique, noté $\overline{X}$ ou μ, est la mesure la plus commune de tendance centrale, elle se définit comme la somme des scores divisée par le nombre de scores. Par exemple, en biologie la moyenne peut être résumée par la somme des observations divisée par l'effectif de l'échantillon étudié:

$$\overline{X} = \frac{\sum X}{N}$$

Elle est calculée pour les caractères quantitatifs.

*** Calcul à partir du tableau élémentaire :**
La moyenne est la somme des valeurs divisée par le nombre d'éléments :

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

*** Calcul à partir du tableau de dénombrement :**
On effectue une moyenne pondérée en assimilant chaque classe j à son centre X_j et en pondérant par l'effectif n_j de la classe.

$$\overline{X} = \frac{\sum_{j=1}^{k} (X_j * n_j)}{N}$$

*** Moyenne pondérée :**

Plus généralement, on recourt à la pondération lorsque les unités n'ont pas le même poids. Si chaque unité i est décrite par sa modalité xi et son poids pi, la moyenne pondérée est :

$$\overline{X}_p = \frac{\sum_{i=1}^{n}(X_i * p_i)}{\sum_{i=1}^{n} p_i}$$

*** Propriétés de la moyenne**

1) Si A = **moyenne de X**

$$\sum_{i=1}^{n} X_i = n * \overline{X}$$

2) La somme des écarts à la moyenne est égale à zéro.

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0$$

3) La moyenne minimise les distances au carré

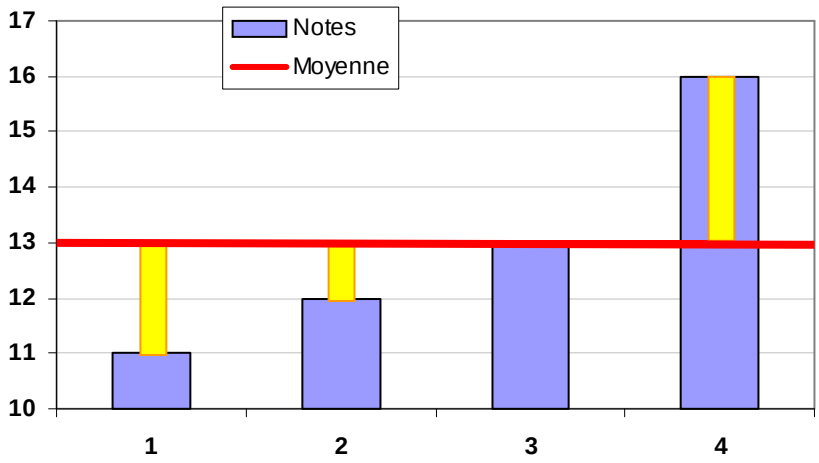
$$\sum_{i=1}^{n} (X_i - A)^2$$

est minimum si ,et seulement si, A est la moyenne du caractère X

Application III.2 :

1) **Exemple avec** illustration
Soit les valeurs de quatre notes : 10, 12, 13 et 16, la moyenne arithmétique est:
(11 + 12 + 13 + 16) / 4 = 13

Illustration



La moyenne arithmétique donne une valeur telle que la somme des écarts (rectangles jaunes) est nulle

La somme de n fois la moyenne donne la somme des n valeurs

Les nombres a, b, c, d, ... sont dits en **progression arithmétique**, dans cet ordre, si la distance qui les sépare est constante : b - a = c -b = d - c = ...

Avec trois nombres, si 2b = a + c alors b est la moyenne arithmétique de a et c

2) Exemple

Soit la série statistique suivante :

valeurs	0	1	2	3	4
effectifs	1	2	1	4	2

La moyenne est : $\overline{x} = \frac{0 + 1 + 1 + 2 + 3 + 3 + 3 + 3 + 4 + 4}{1 + 2 + 1 + 4 + 2} = \frac{24}{10} = 2{,}4$

On préférera écrire : $\overline{x} = \frac{0 + 2\times1 + 2 + 4\times3 + 2\times4}{1 + 2 + 1 + 4 + 2} = \frac{24}{10} = 2{,}4$

3) Calcul de la moyenne

Soit la série statistique suivante :

valeurs	x_1	x_2	…	x_p
effectifs	n_1	n_2	…	n_p

La <u>moyenne</u> est : $\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{n_1 + n_2 + \ldots + n_p}$

Remarque :

Si les données ont été regroupées en classes, on ne peut calculer la valeur exacte de la moyenne. On peut toutefois en déterminer une bonne approximation en remplaçant chaque classe par son milieu.

4) Dans les séries statistiques suivantes déterminer les moyennes :

a) Tableau de fréquences

valeurs	12	13	14	15	16
fréquences	0,05	0,17	0,43	0,30	0,05

$\overline{x}$ =

b) Données réparties en classes

classes	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20]
effectifs	7	12	14	2

Remplaçons chaque classe par son milieu :

$\overline{x} \approx$

5) Propriétés

a) Addition ou Multiplication de toutes les données par un même nombre :

Ex : Soit la série : 10, 12, 14. $\overline{x} =$
Ajoutons 2 : la nouvelle série est : 12, 14, 16. $\overline{x} =$
Divisons par 2 : la nouvelle série est : 6, 7, 8. $\overline{x} =$

Cas général : Soit α un réel quelconque :
- Si l'on ajoute α à toutes les données, la moyenne augmente d'α
- Si on multiplie toutes les données par α, la moyenne est multipliée par α
- Si on divise toutes les données par α, la moyenne est divisée par α

b) Moyennes partielles

Ex : Sur un patient diabétique après 10 prises de sang, le taux moyen de glycémie est régulé à 1,25g/l. La valeur de la glycémie à la $11^{ème}$ prise est de 0,8 g/l. Quel est le nouveau taux moyen de glycémie de ce patient ?
- Calculons la somme des 10 prises de sang = 1,25 x 10 = 12,5 g/l
- Calculons la nouvelle somme des 11 prises de sang = 1,25 x 10 + 0,8 = 13,3 g/l
- Calculons la nouvelle moyenne des 11 prises de sang = 13,3/11 = 1,20g/l

Cas général : Si on réuni deux groupes disjoints ayant respectivement pour moyennes et effectifs, $\overline{x}_1$ et n_1 d'une part, $\overline{x}_2$ et n_2 d'autre part, la moyenne de l'ensemble sera alors :

$$\overline{x} = \frac{n_1 \times \overline{x}_1 + n_2 \times \overline{x}_2}{n_1 + n_2}$$

Application III.3 : Calculer des moyennes en utilisant le logiciel Excel et calculer la moyenne de l'exemple 2 de l'Application III.2 :

6.1.2. 1. Calcul de la moyenne par changement d'origine et d'unité.
(Voir partie Application VII : Approfondissement)

6.1.2.2. Autres indicateurs de moyenne :
Il existe des indicateurs de la moyenne autre que la moyenne arithmétique. Néanmoins, ils sont moins utilisés en biostatistique car ils ne présentent d'intérêt que dans des cas très particuliers. Ils ne feront pas l'objet de ces modules: la moyenne géométrique, la moyenne harmonique, la moyenne quadratique, la moyenne arithmético-géométrique.

6.1.3. La médiane et la classe médiane

6.1.3.1. Définition générale :

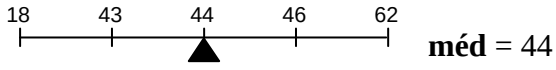
On appelle **<u>médiane</u>** la valeur "du milieu". On dit qu'elle partage la série statistique en deux moitiés : il y a autant de valeurs en dessous qu'au dessus. (C'est la donnée qui permet de diviser une série ordonnée d'une façon croissante en 2 parties égales (50%, 50%). La médiane ne peut être calculée que pour les caractères quantitatifs.

6.1.3.2. Médiane, pour les données rangées : Les valeurs du caractère X étant classées par ordre croissant, la médiane est la valeur du caractère qui partage l'ensemble décrit par X en deux sous ensembles d'effectifs égaux : 50 % des éléments ont des valeurs de X supérieures à $X_{méd}$ et 50% prennent des valeurs inférieures.

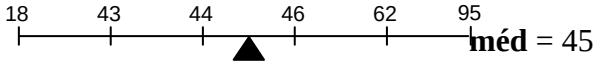
- Méthode

Soit une série statistique d'effectif total *n*, rangée par ordre croissant.
Pour déterminer son rang, il y a 2 cas :

- si *n* est impair : la médiane est la valeur de rang $\frac{n+1}{2}$



- si *n* est pair : nous prendrons la demi-somme des deux valeurs dont les rangs entourent le nombre $\frac{n+1}{2}$



<u>APPLICATION IV</u>

<u>Application IV.1</u>

Cas de données discrètes "en vrac" ***10, 7, 12, 18, 16, 15, 5, 11, 11, 20, 15, 11, 18, 14***

Ordonnons la série par ordre croissant : 5, 7, 10, 11, 11, 11, 12, 14, 15, 15, 16, 18, 18, 20

Il y a 14 termes or la valeur de rang est $\frac{14+1}{2} = 7{,}5$.

La médiane est donc la demi somme des $7^{ème}$ et $8^{ème}$ termes : médiane = $\frac{12+14}{2} = 13$

6.1.3.3. Médiane, pour les données condensées : La définition est la même, elle correspond dans ce cas à la première modalité ou valeur dont la fréquence relative cumulée dépasse 0,500 ou l'effectif cumulé dépasse les 50%.

Méthode :

Il faut calculer les fréquences ou les effectifs cumulés dès que celle-ci atteint respectivement 0.5 ou 50% il suffit de choisir le nombre à mi chemin entre la modalité ou valeur concernée et la suivante.

Application IV.2

Cas d'un tableau d'effectifs

On ordonne le tableau, et on cherche l'élément qui partage la distribution en deux parties égales: on repère l'élément qui a le rang (N+1)/2 pour le caractère X. Si la distribution a un nombre impair d'éléments on trouve une valeur unique qui est la médiane, si la distribution a un nombre pair d'éléments, on trouve deux valeurs qui déterminent un **intervalle médian** : on prend alors pour médiane le centre de cet intervalle médian.

valeurs	1	2	**3**	4	5	6
effectifs	6	11	**25**	19	15	5
effectifs cumulés	6	17	**42**	61	76	81
fréquence		17	**42**	61	76	81
fréquence	0,07	0,14	**0,31**	0,23	0,19	0,06
fréquences cumulées	0,07	0,21	**0,52**	0,75	0,94	1,00

Attention, il faut bien interpréter cette dernière ligne : Les données qui valent 3 ont un rang compris entre 18 et 42 inclus

L'effectif total est de 81 or la valeur de rang $\frac{81+1}{2} = 41$.

La médiane est donc le $41^{ème}$ terme : médiane = 3

6.1.3.4. Médiane, pour les données réparties par classes

Remarque

Si les données ont été regroupées en classes, on ne peut déterminer la valeur exacte de la médiane. En revanche, on appellera **classe médiane**, la classe qui la contient (et permet donc d'en donner un encadrement).

La classe médiane est la première classe où la fréquence cumulée est supérieure à 0,500.

Application IV.3

classe	[0 ; 2[	[2 ; 4[	[4 ; 6[	[6 ; 8]
fréquence	10%	38%	45%	7%
fréquence cumulée	10%	48%	93%	100%

48% des valeurs sont strictement inférieures à 4

Et 93% des valeurs sont strictement inférieures à 6

La classe médiane est donc la classe [4 ; 6[

On peut donc en déduire l'encadrement suivant $4 < méd < 6$

Méthode de calcul

Pour préciser la valeur de la médiane, il faut supposer que toutes les données sont réparties uniformément (c'est-à-dire que les données sont réparties sur un continuum).

On repère la classe qui contient la médiane, puis on réalise une interpolation linéaire pour estimer la valeur de celle-ci selon la formule suivante :

$$Md = Bmd + (\frac{0{,}500 - Fmd_{-1})}{Fmd} Lmd$$

Où : Bmd : Borne inférieure de la classe médiane

Fmd-1 : Fréquence relative cumulée de la classe qui précède la classe médiane.

Fmd : Fréquence relative de la classe médiane.

Lmd : largeur, amplitude des classes

Application pour l'exemple précédent :

$$Md = 4 + (\frac{0{,}500 - 0{,}48)}{0{,}45} 2 = 4{,}088$$

Remarques :

Rq1 : Autre méthode de calcul de la médiane : il est aussi possible de déterminer la médiane à l'aide des polygones des effectifs cumulés (voir TD)

Rq2 : Propriétés de la médiane : La médiane est la valeur du caractère qui est la plus proche de toutes les autres. C'est celle qui minimise les distances en valeur absolue :

$\sum_{i=1}^{N} |x_i - x_{méd}|$ **est minimum si et seulement si** $x_{méd}$ **est la médiane du caractère X**

6.1.4 Quantiles : Mesures de position statistique en référence à la médiane

Il a été vu précédemment que la médiane partage la distribution des fréquences en 2 parties égales. Il est possible de partager une distribution de fréquence en 4 parties égales (quartiles), en 10 parties égales (déciles), en 100 parties égales (centiles), en n parties égales….

6.1.4.1. Définition des quantiles : on appelle quantiles les valeurs du caractère qui définissent les bornes d'une partition en classes d'effectifs égaux.

6.1.4.2. Les quartiles sont les trois valeurs qui permettent de découper la distribution en quatre classes d'effectifs égaux. On les notes X_{Q1}, X_{Q2} et X_{Q3}

Représentation des quartiles

Partition du caractère	X_{min} intervalle interquartile 1	$\mathbf{X_{Q1}}$ ¼ Quartile inférieur	intervalle interquartile 2	$\mathbf{X_{Q2}}$ ½ Médiane	intervalle interquartile 3	$\mathbf{X_{Q3}}$ ¾ Quartile supérieur	X_{max} intervalle interquartile 4
Fréquence des éléments	**25%**		**25%**		**25%**		**25%**

- Q1 : quartile inférieur, 25% des valeurs de la variable lui sont inférieures et 75% lui sont supérieures
- Q2 : médiane, 50% des valeurs de la variable lui sont inférieures et 50% lui sont supérieures
- Q1 : quartile supérieur, 75% des valeurs de la variable lui sont inférieures et 25% lui sont supérieures

Remarque : X_{Q2} est égal à la médiane.

6.1.4.3. Les **déciles** sont les 9 valeurs de X qui permettent de découper la distribution en dix classes d'effectifs égaux. 0n les note $X_{d1}...X_{d9}$.

Représentation des déciles

Partition du caractère	X_{min} **Int-1**	$\mathbf{X_{d1}}$ 1/10	**Int-2**	$\mathbf{X_{d2}}$ 1/20	**Int-3**	$\mathbf{X_{d3}}$ 1/30	…..	$\mathbf{X_{d8}}$ 1/20	**Int-9**	$\mathbf{X_{d9}}$ 9/10	X_{max} **Int-10**
Fréquence des éléments	**10%**		**10%**		**10%**				**10%**		**10%**

Int-(intervalle interdécile)

- **D1 :** décile inférieur, 10% des valeurs de la variable lui sont inférieures et 90% lui sont supérieures
- **D2 :** 20% des valeurs de la variable lui sont inférieures et 80% lui sont supérieures
- **D3 :** 30% des valeurs de la variable lui sont inférieures et 70% lui sont supérieures
- D4 :….
- **D5 :** médiane, 50% des valeurs de la variable lui sont inférieures et 50% lui sont supérieures

….

- **D9 :** décile supérieur, 90% des valeurs de la variable lui sont inférieures et 10% lui sont supérieures

6.1.4.4. Les **centiles** sont les 99 valeurs de X qui permettent de découper la distribution en 100 classes d'effectifs égaux. 0n les note $X_{c1}...X_{c99}$.

Remarques
Les différentes mesures de position (quartile, décile,…) ne sont que des cas particuliers des centiles.
Les centiles sont donc très utiles pour déterminer les valeurs des autres mesures de positions

Exemple de correspondances entre mesures de positions

Q1= C25 = 25ème centile
Q2= C50 = D50 = 50ème centile = Médiane
Q3 = C75 = 75ème centile
D1= C10 = 10ème centile
D2= C20 = 20ème centile
…
D9 = C90 = 90ème centile

6.1.4.5. Calculs des quantiles
Nous nous limiterons aux cas des centiles car nous pouvons facilement faire des correspondances avec les autres mesures de positions.

6.1.4.5.1. Détermination des valeurs de la variable à partir d'un rang centile données.
$C\alpha$: rang du centile (le rang est donnée, quelle est la valeur de la variable correspondant à ce rang ?)

a) Cas des données rangées :

$C\alpha$: rang du centile : Il correspond à la donnée dont le rang est l'entier qui suit : $\frac{N\alpha}{100}$ si $\frac{N\alpha}{100}$ n'est pas un entier. Dans le cas contraire si $\frac{N\alpha}{100}$ est un entier, **$C\alpha$** correspond à la données dont la position (le rang) est a mis chemin entre le rang donnée par : $\frac{N\alpha}{100}$ et la position suivante

N : nombre total de valeurs dans la série statistique
α : le rang du centile

Application IV.4

Exemples :
Soit la série statistique suivante :
58 ; 59 ; 64 ; 64 ; 64 ; 68 ; 71 ; 71 ; 79 ; 82 ; 82 ; 85 ; 92 ; 92 ; 92 ; 95
- trouver les centiles suivants : C15 ; C40
- trouver les quartiles : Q2 et Q3

Réponses :
N : 16

- Pour centile C15 :

α =15, le rang de la donnée est déterminé par la formule $\frac{N\alpha}{100} = \frac{16x15}{100} = 2{,}4$

La valeur n'est pas un entier, le rang est donc le premier entier suivant 2,4 ainsi C15 correspond au rang 3, ce dernier correspond à la valeur : 64

- Pour centile C40 (qui correspond au décile 4) :

α =20 le rang de la donnée est déterminé par la formule $\frac{N\alpha}{100} = \frac{16x40}{100} = 6{,}4$

La valeur n'est pas un entier, le rang est donc le premier entier suivant 6,4 ainsi C40 (ou D4) correspond au rang 7, ce dernier correspond à la valeur : 71

- Pour centile C50 ou quartile Q2 ou la médiane (qui correspond au décile 5) :

α =50 le rang de la donnée est déterminé par la formule $\frac{N\alpha}{100} = \frac{16x50}{100} = 8$

La valeur est un entier, **Cα** correspond à la données dont la position (le rang) est a mis chemin entre le rang 8 et le rang 9, ainsi Q2 correspond à la moyenne des valeurs du au rang 8 (qui correspond à la valeur 71) et le rang 9 (qui correspond à la valeur 71) : $Q2 = \frac{71+79}{2} = 75$

- Pour centile C75 ou quartile 3:

α =75 le rang de la donnée est déterminé par la formule $\frac{N\alpha}{100} = \frac{16x75}{100} = 12$

La valeur est un entier, **Cα** correspond à la données dont la position (le rang) est a mis chemin entre le rang 12 et le rang 13, ainsi Q3 correspond à la moyenne des valeurs du au rang 12 (qui correspond à la valeur 85) et le rang 13 (qui correspond à la valeur 92) : $Q3 = \frac{85+92}{2} = 88{,}5$

b) Cas des données condensées :
La méthode est identique à la précédente, mais il est aussi possible de travailler avec les fréquences relatives. Dans le cas de détermination avec les fréquences, Cα correspond à la première modalité dont la fréquence cumulée dépasse $\frac{\alpha}{100}$. Dans le cas où $\frac{\alpha}{100}$ est un entier, il suffira de choisir le nombre à mi-chemin entre la modalité concernée et la suivante.

Application IV.5

Calculons C69 pour des données condensées

xi	ni	eff cum	fi	Fi (freq cum)
128	8	8	0,11	0,11
145	13	21	0,18	0,29
160	14	35	0,19	0,49
180	**16**	**51**	**0,22**	**0,71**
195	11	62	0,15	0,86
197	7	69	0,10	0,96
209	3	72	0,04	1,00
Somme	**72**			

Choisir C = 69
calcul avec les effectifs
C 69 = 49,68
calcul avec les fréquences
C 69 = 0,69

Pour le calcul avec les effectifs : la formule est la suivante : (N=72)

$$\frac{N\alpha}{100} = \frac{72x69}{100} = 49{,}68$$

C69 correspond à la modalité occupant le rang 50 dans la distribution, elle correspond donc à la valeur 180

Pour le calcul avec les fréquences : la formule est la suivante :

$$\frac{\alpha}{100} = 69/100 = 0{,}69$$

C69 correspond à la modalité dont la fréquence relative cumulée dépasse 0,69. Dans la distribution, cette fréquence correspond à la valeur 180

c) Cas des données groupées en classes :

La classe contenant Cα correspond à la première classe où la fréquence cumulée atteint ou dépasse $\frac{\alpha}{100}$, par référence à la formule du calcul de la médiane (vue précédemment) il est possible d'écrire la formule suivante de Cα

$$C\alpha = Bc\alpha + (\frac{\frac{\alpha}{100} - Fc\alpha_{-1}}{Fc\alpha})Lc\alpha$$

Où : Bcα : Borne inférieure de la classe contenant cα
Fcα -1 : Fréquence relative cumulé de la classe qui précède la classe contenant cα
Fcα : Fréquence relative de la classe contenant cα.
Lcα : largeur, amplitude de la classe contenant cα

Application IV.6

Calculer C80 des classes suivantes :

limites inférieures des classes (cm)	mi	ni	eff cum	fi	Fi (freq cum)
130	135	12	12	0,12903	0,1290
140	145	20	32	0,21505	0,344
150	155	24	56	0,25806	0,602
160	**165**	**21**	**77**	**0,22581**	**0,828**
170	175	11	88	0,11828	0,946
180	185	5	93	0,05376	1,000
Somme		**93**		1,00000	

La classe contenant Cα (C80) est la première classe où Fi > $\frac{\alpha}{100} = 80/100 = 0{,}80$

C80 correspond à la classe [160-170[

Calcul de la valeur de la modalité correspondant à C80

$$C\alpha = Bc\alpha + (\frac{\frac{\alpha}{100} - Fc\alpha_{-1}}{Fc\alpha})Lc\alpha$$

Bcα : Borne inférieure de la classe contenant C80 = **160 cm**
Fcα -1 : Fréquence relative cumulé de la classe qui précède la classe contenant C80 = **0,828**
Fcα : Fréquence relative de la classe contenant C80 = **0,22581**
Lcα : largeur, amplitude de la classe contenant C80 = 130-140=150-140=…=**10cm**

AN (application numérique)

$$C80 = 160 + (\frac{\frac{80}{100} - 0{,}828}{0{,}22581})10 = 168{,}7619cm$$

6.1.4.5.2. Détermination du rang centile à partir d'une valeur donnée de la variable.
Cet détermination est le processus inverse par rapport aux éléments précédent du paragraphe : 5.1.4.5.1., ce qui consiste à recherche Cα pour une valeur connu Xi d'une série statistique X.

a) Cas des données rangées ou condensées
Il suffit de calculer simplement le pourcentage des données inférieures à la valeur (ou observation) donnée.

Application IV.7

Exemple 1 : série ordonnée croissante

Dans les valeurs de la glycémie de la série statistique suivante trouver le centile Cα de la valeur 0,96g/l :

0,6 g/l; 0,6 g/l; 0,65 g/l; 0,7 g/l; 0,72 g/l; 0,72 g/l; 0,72 g/l; 0,74 g/l; 0,75 g/l; 0,75 g/l; 0,76 g/l; 0,78 g/l; 0,78 g/l; 0,8 g/l; 0,8 g/l; 0,83 g/l; 0,83 g/l; 0,84 g/l; 0,84 g/l; 0,84 g/l; 0,9 g/l; 0,96 g/l; 1,01 g/l; 1,02 g/l; 1,1 g/l; 1, 15 g/l; 1,16 g/l; 1,18g/l ; 1,2g/l.

Il s'agit de trouver le pourcentage des données dont la valeur de la glycémie est inférieure à 0,96g/l : Cette valeur est à la 22 positions (22ème valeur de la série ordonnée de façon croissante), il y a 21 valeurs de la glycémie inférieures à 0,96g/l sur un total de 29 valeurs (N= 29),le pourcentage est donc de : $100 \text{ X } (\frac{21}{29}) = 72{,}41\%$, ainsi **le rang centile Cα de la valeur de la glycémie de 0,96g/l est de 72** (la valeur de 0,96g/l de glycémie correspond au centile C72)

Application IV.8

Exemple 2: tableau de distribution condensée

Dans le tableau de distribution des valeurs de la glycémie suivante trouver le centile Cα de la valeur 1,1g/l :

xi (g/l)	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1	1,05	1,1	1,15	1,2
ni	8	12	24	26	32	32	28	26	21	24	**20**	18	11
ni (cum)	8	20	44	70	102	134	162	188	209	**233**	**253**	271	282

Il s'agit de trouver le pourcentage des données dont la valeur de la glycémie est inférieure à 1,1g/l : Cette valeur est à la 253 positions (253 ème valeurs des effectifs cumulés), il y a 233 valeurs de la glycémie inférieures à 1,1g/l sur un total de 282 valeurs (N= 282), le pourcentage est donc de : $100 \text{ X } (\frac{233}{282}) = 82{,}62\%$, ainsi **le rang centile Cα de la valeur de la glycémie de 1,1g/l est de 82** (au moins 82% des valeurs de la glycémie sont inférieures à 1,1g/l).

b) Cas des données rangées en classes

Le rang centile Cα d'une donnée (ou observation) est obtenu par la formule suivante

$$100\left[\left(\frac{x_r - b_r}{L_r}\right) fr + F_{r-1}\right]$$

Xr : la donnée dont on recherche le rang centile Cα
br : borne inférieure de la classe contenant Xr
Lr : largeur de la classe contenant Xr
fr : fréquence relative de la classe contenant Xr
Fr-1 : fréquence relative cumulée de la classe qui précède la classe contenant Xr

Application IV.9

Exemple : tableau de distribution groupé en classes

Dans le tableau de distribution des valeurs de la glycémie suivante trouver le centile Cα de la valeur 1,1g/l :

xi (g/l)	[0,6-0,7[	[0,7-0,8[	[0,8-0,9[	[0,9-1,0[	[1,0-1,1[	[1,1-1,2[	[1,2-1,3[	[1,3-1,4[	[1,4-1,5[	somme
ni	20	18	26	28	29	25	21	20	21	**208,00**
fi	0,10	0,09	0,13	0,13	0,14	0,12	0,10	0,10	0,10	**1,00**
ni (cum)	20	38	64	92	121	146	167	187	208	
Fi cum	0,10	**0,18**	0,31	0,44	0,58	0,70	0,80	0,90	1,00	

Recherchons les rangs centile des valeurs de la glycémie suivante :
0,81g/l ; 1,12g/l et 1,18g/l

*** Pour la valeur de la glycémie de 0,8g/l**

0,8g/l se situe dans la classe [0,8-0,9[, le rang centile de 0,8g/l est l'entier inférieur à :

$$100\left[\left(\frac{x_r - b_r}{L_r}\right) fr + F_{r-1}\right]$$

Xr : la donnée dont on recherche le rang centile Cα =0,81
br : borne inférieure de la classe contenant Xr = 0,80
Lr : largeur de la classe contenant Xr = 0,1
fr : fréquence relative de la classe contenant Xr = 0,13
Fr-1 : fréquence relative cumulée de la classe qui précède la classe contenant Xr=0,18

Application numérique

$$100\left[\left(\frac{0,81-0,8}{0,1}\right)0,13+0,18\right]=19,3$$

Le rang centile de 0,81g/l est 19, ainsi au moins 19% des données sont inférieures à 0,81g/l

*** Pour les valeurs de la glycémie de 1,12g/l et 1,18g/l**

Elles sont les 2 se situées dans la classe [1,1-1,2[, le rang centile de **1,12g/l et 1,18g/l** est l'entier inférieur à :

$$100\left[\left(\frac{x_r - b_r}{L_r}\right) fr + F_{r-1}\right]$$

Xr : la donnée dont on recherche le rang centile Cα =1,12
br : borne inférieure de la classe contenant Xr = 1,1
Lr : largeur de la classe contenant Xr = 0,1
fr : fréquence relative de la classe contenant Xr = 0,12
Fr-1 : fréquence relative cumulée de la classe qui précède la classe contenant Xr=0,58

Application numérique

$$100\left[\left(\frac{1,12-1,1}{0,1}\right)0,12+0,58\right]=60,4$$

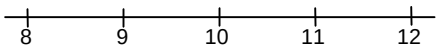
Le rang centile de 1,12/l est 60, ainsi au moins 60% des données sont inférieures à 1,12/l

Le rang centile de 1,18/l est 67, ainsi au moins 67% des données sont inférieures à 1,18/l

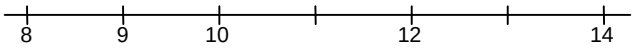
Calcul de Cα	**Valeurs**	**Valeurs**	**Valeurs**
Xr : la donnée dont on recherche le rang centile Cα	**0,81**	**1,12**	**1,18**
br : borne inférieure de la classe contenant Xr	**0,8**	**1,1**	**1,1**
Lr : largeur de la classe contenant Xr	**0,1**	**0,1**	**0,1**
fr : fréquence relative de la classe contenant Xr	**0,13**	**0,12**	**0,12**
Fr-1 : fréquence relative cumulée de la classe qui précède la classe contenant Xr	**0,18**	**0,58**	**0,58**
xr-br	**0,01**	**0,02**	**0,08**
(xr-br)/Lr	**0,1**	**0,2**	**0,8**
((xr-br)/Lr)xfr	**0,013**	**0,024**	**0,096**
(((xr-br)/Lr)xfr)+Fr-1	**0,193**	**0,604**	**0,676**
(((xr-br)/Lr)+Fr-1)x100	**19,3**	**60,4**	**67,6**
Cα	**19**	**60**	**67**

6.1.5. Moyenne et médiane

- Quand on modifie les valeurs extrêmes d'une série, la moyenne change contrairement à la médiane qui ne change pas. On dit que la moyenne est "sensible aux valeurs extrêmes". Il arrive que certaines de ces valeurs extrêmes soient douteuses ou influent de façon exagérée sur la moyenne. On peut alors, soit calculer une moyenne élaguée (c'est à dire recalculer la moyenne sans ces valeurs gênantes), soit utiliser la médiane.
- Comment interpréter un écart entre la moyenne et la médiane ?

Soit la série suivante : 

Ici la moyenne et la médiane sont identiques : la série est bien "centrée".

Soit la nouvelle série : 

Ici la moyenne est plus importante que la médiane : la série est plus "étalée à droite".

6.1.6. Avantages et inconvénients des différentes valeurs centrales :

Le statisticien Yule (XIXème siècle) a défini six propriétés souhaitables pour les valeurs centrales. Le tableau ci-dessous permet de montrer les avantages et inconvénients des trois valeurs centrales (Mode, Médiane, Moyenne arithmétique)

Propriétés	**Mode**	**Médiane**	**Moyenne**
1) est définie de façon objective	+	+	+
2) dépend de toutes les valeurs observées	-	-	+
3) a une signification concrète	+	+	-
4) est simple à calculer	+	+	+
5) est peu sensible aux fluctuations de l'échantillon	-	+	-
6) se prête au calcul algébrique	-	-	+

***Tableau 3** : Avantages et inconvénients des trois valeurs centrales (propriété + réalisée, - non réalisée)*

6.2. Paramètre de dispersion

Dispersion statistique : On appelle dispersion statistique, la tendance qu'ont les valeurs de la distribution d'un caractère à s'étaler, à se disperser, de part et d'autre d'une valeur centrale. On

distingue la dispersion absolue (mesurée dans l'unité de mesure du caractère) et la dispersion relative (mesurée par un nombre sans dimension).

6.2.1. Les paramètres de dispersion absolue

Les paramètres de dispersion absolue indiquent de combien les valeurs d'une distribution s'écartent en général de la valeur centrale de référence. Un paramètre de dispersion absolue s'exprime toujours dans l'unité de mesure de la variable considérée. Les quatre paramètres de dispersion absolue les plus courants sont :
- l'étendue,
- l'intervalle inter quantile (écarts inter quantiles),
- l'écart absolu moyen
- l'écart type.

6.2.1.1 L'étendue de la variation: l'étendue d'une distribution est égale à la différence entre la plus grande et la plus petite valeur de la distribution :

Etendue de X = Xmax - Xmin

Plus l'étendu est grande plus les valeurs sont dispersées.

Exemple : l'étendue est donnée par la valeur minimale et la valeur maximale : dans le cas de l'exemple précédent il s'agit de la différence : 14 mm - 4 mm = 10 mm
• La moyenne et la médiane sont les estimateurs statistiques du centre d'une distribution

a) cas de données rangées :

L'étendu de la distribution de la série statistique :
0,5 g/l; 0,58 g/l; 0,65 g/l; 0,7 g/l; 0,72 g/l; …….; 1, 15 g/l; 1,16 g/l; 1,18g/l ; 1,2g/l.
La plus grande valeur est: 1,2g/l
La plus petite valeur est :0,6g/l
L'étendue de la variation : 1,2-0,5 = 0,7

b) cas de données groupées en classes :

[0,6-0,7[	[0,7-0,8[	[0,8-0,9[	[0,9-1,0[	[1,0-1,1[	[1,1-1,2[	[1,2-1,3[	[1,3-1,4[	[1,4-1,5[

La dernière classe a comme borne supérieure = 1,5
La première classe a comme borne inférieure =0,6
L'étendue de la variation est : 1,5 – 0,6 = 0,9

6.2.1.2. Quartile et intervalle interquartile : Mesures de la dispersion statistique en référence à la médiane

6.2.1.2.1. L'intervalle interquartile est l'étendue de la distribution sur laquelle se trouvent concentrée la moitié des éléments dont les valeurs de X sont les plus proches de la médiane. On exclut alors de la distribution les 25% des valeurs les plus faibles et les 25 % des valeurs les plus fortes de X. Cet intervalle se note:(X_{q3}-X_{q1}).

6.2.1.2.2. L'intervalle inter-décile est l'étendue de la distribution sur laquelle se trouvent concentrés 80% des éléments dont les valeurs de X sont les moins différentes de la médiane. On exclut alors de la distribution les 10 % des valeurs les plus faibles et les 10% des valeurs les plus fortes. Il se note (X_{d9}-X_{d1}).

6.2.1.3. Mesures de la dispersion statistique en utilisant l'écart semi-interquartile
Cet écart mesure la moitié de l'étendue de la moitié centrale des données
Il est calculé selon la formule suivante :

$$Q = \frac{Q_3 - Q_1}{2} = \frac{Q_{75} - Q_{25}}{2}$$

APPLICATION V :

Application V.1 (calculons l'écart semi-interquartile)

- Cas 1 : Données rangées
Le tableau de distribution obtenu par Excel

Rang	1	2	3	4	5	6	7	8	9	10
Variables	**20**	**21**	**31**	**33**	**34**	**36**	**36**	**39**	**40**	**43**
	N =	10	**Moyenne =**	33,3	**Mode =**	36	**Q1** =	31,5	**Q3** =	38,25
					Médiane = Q2	35	**Q1 arrondi =**	32	**Q3 arrondi =**	39
							Semi – Interquartile =		**3,5=4**	
							29,8 **< 50% des valeurs <** 36,8			

Rappel Qn correspond à la donnée dont le rang est l'entier qui suit la formule

Cα : rang du centile : Il correspond à la donnée dont le rang est l'entier qui suit : $\frac{N\alpha}{100}$

Q_1 et Q_3 correspondent respectivement au C_{25} et C_{75}

Calculs :
N= 10 (nombre de données pour le caractère étudié), la **Moyenne = 33,3**

- Pour rechercher la donnée correspondant au **Q1, (au centile 25**), il suffit de calculer $\frac{N\alpha}{100}$

$\frac{10x25}{100} = 2,5$ de prendre l'approximation supérieure ce qui correspond au 3ème rang et donc à la donnée 31

- Pour rechercher la donnée correspondant au **Q3, (au centile 75**), il suffit de calculer $\frac{N\alpha}{100}$

$\frac{10x75}{100} = 7,5$ de prendre l'approximation supérieure ce qui correspond au 8ème rang et donc à la donnée 39

Donc : $Q = \frac{Q_3 - Q_1}{2} = \frac{Q_{75} - Q_{25}}{2} = \frac{39 - 31}{2}$ **= 4 (sur Excel = 3,5)**

Interprétation : Contenu de la valeur de la moyenne = 33,3 il y statistiquement 50% des valeurs de la série numérique comprise entre 29,8 et 36,8
Moyenne – Q < 50% des valeurs < et Moyenne + Q

- Cas 2 : Données condensées :

Variables	**40**	**45**	**48**	**52**	**56**	**58**	**66**	**70**	**Total**
ni (effectif)	8	6	12	**24**	26	28	15	8	**127**
ni Cumulé	**8**	**14**	**26**	**50**	**76**	**104**	**119**	**127**	
nixi	320	270	576	1248	1456	1624	990	560	7044
Total eff	127								
Moyenne	55,46								
Mode	8								

Calcul des Quartiles (par méthode des centiles)					
	Centile	Rang	Rang Arrondi >	Quartiles	Valeurs
Q1 = Cn =>	25	31,75	32	**Q1 =**	52
Q3 = Cn =>	75	95,25	96	**Q3 =**	58
Q2 = Cn =>	50	63,5		**Q2 = Médiane =**	56
Semi – Interquartile =		**3**			
	52,46	**< 50% des valeurs <**		58,46	

Calculs :
N= 127 (nombre de données pour le caractère étudié),
N= 8 + 6 + 12 +…+8 = 127, la **Moyenne = 7044/127 =55**

- Pour rechercher la donnée correspondant au **Q1, (au centile 25**), il suffit de calculer $\frac{N\alpha}{100}$

$\frac{127x25}{100} = 31{,}75$ de prendre l'approximation supérieure ce qui correspond à la valeur cumulée 50, (il suffit de choisir dans les effectifs cumulés la valeur qui est supérieure à 31,75 ce qui correspond à la valeur cumulée 50) puis par correspondance déterminer la variable qui correspond à cette valeur cumulée. Ainsi la valeur cumulée 50 correspond à la variable 52 => Q1 (centile 25) =52

- Pour rechercher la donnée correspondant au **Q3, (au centile 75**), il suffit de calculer :

$\frac{N\alpha}{100} = \frac{127x75}{100} = 95{,}25$

Puis, de prendre l'approximation supérieure, ce qui correspond à la valeur cumulée 104, (il suffit de choisir dans les effectifs cumulés la valeur qui est supérieure à 95,25 ce qui correspond à la valeur cumulée 104) puis par correspondance déterminer la variable qui correspond à cette valeur cumulée. Ainsi, la valeur cumulée 104 correspond à la variable 58 => Q3 (centile 75) = 58

Donc : $Q = \frac{Q_3 - Q_1}{2} = \frac{Q_{75} - Q_{25}}{2} = \frac{58 - 52}{2}$ **= 3 (sur Excel = 3)**

Interprétation : Compte tenu de la valeur de la moyenne = 55, il y statistiquement 50% des valeurs de la série numérique comprises entre 52 et 58
Moyenne – Q < 50% des valeurs < et Moyenne + Q

- Cas 3 : Données groupés en classes :
Le calcul de l'écart semi-interquartile sera traité par chaque étudiant avec le logiciel Excel. Pour cela, utiliser les valeurs du tableau de distribution des valeurs de la glycémie de l' Application IV.9

6.2.1.4. Mesures de la dispersion statistique en référence à la moyenne arithmétique

6.2.1.4.1. Ecart absolu moyen ou Ecart Moyen Absolu « EMA »: Ce paramètre est la moyenne arithmétique de la valeur absolue des écarts à la moyenne. Il correspond à la moyenne des valeurs absolues de chaque donnée par rapport à la moyenne.

a) Données rangées :

L'écart absolu moyen est la moyenne des distances mesurées positivement (en valeur absolue) entre les données et la moyenne.

$$EMx = \frac{\sum_{i=1}^{n} |xi - x|}{N}$$

Exemple :

Poids (kg)	65	66	67	68	68	69	70	70	71	71	71	72	73	74	74	75	75	75

N= 18 ; Moyenne = 70,77kg

$$EMA = \frac{|65 - 70{,}77| + |66 - 70{,}77| + |67 - 70{,}77| ..... + |75 - 70{,}77|}{18} = 0{,}143$$

L'écart absolu moyen est faible et les valeurs sont très concentrées autour de la moyenne

b) Données rangées : le calcul de EMA s'exprime par les formules suivantes :

$$EMx = \frac{\sum_{i=1}^{n} ni\,|xi - x|}{N}$$ ce qui équivaut avec les fréquences à la formule $$EMx = \sum_{i=1}^{n} fi\,|xi - x|$$

b) Données groupées en classes : le calcul de EMA s'exprime par l'une des 4 formules suivantes :

$$EMx = \frac{\sum_{i=1}^{n} ni\,|xi - x|}{N}$$ ou $$EMx = \frac{\sum_{i=1}^{n} ni\,|mi - x|}{N}$$

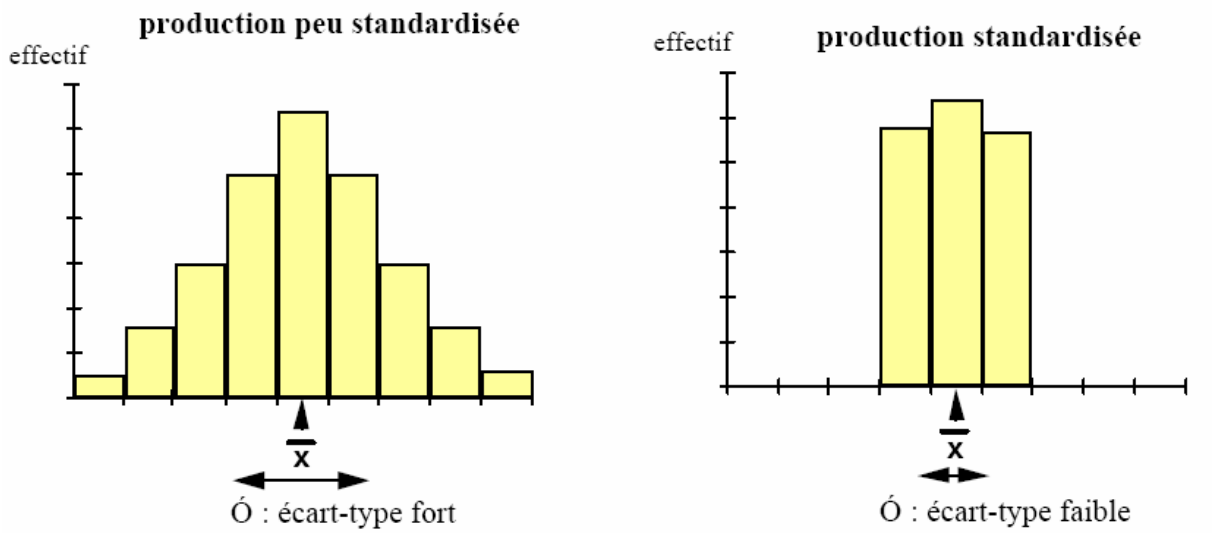
$$EMx = \sum_{i=1}^{n} fi\,|xi - x|$$ ou $$EMx = \sum_{i=1}^{n} fi\,|mi - x|$$

6.2.1.2.2. Variance et écart-type :

La variance et écart-type servent à évaluer la dispersion d'une distribution autour d'une valeur centrale, la moyenne.

Application V.2 : (distribution standardisée ou non ?)

Soit deux séries de microscopes produits dans deux usines différentes. Nous désirons juger de la standardisation de chacune des deux séries. Je choisis de comparer le poids maximal de chaque microscope.
- si les écarts à la moyenne sont faibles la production est standardisée
- si les écarts à la moyenne sont élevés, la production est peu standardisée



a - Variance : La variance, notée $(\sigma_x)^2$ est la moyenne du carré des écarts à la moyenne.

$$(\delta_x)^2 = \frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{N}$$

La variance n'est pas un paramètre de dispersion absolue mais plutôt une mesure globale de la variation d'un caractère de part et d'autre de la moyenne arithmétique (quantité d'information). Pour obtenir un paramètre de dispersion absolue, on effectue la racine carrée de la variance, appelé **écart-type** et que l'on note σx

La variance pour des données rangées ou groupées en classe devient :

$$(\delta_x)^2 = \frac{\sum_{i=1}^{n} ni\ (xi - \bar{x})^2}{N}$$

Ou $\mathbf{n_i}$ désigne les effectifs de chaque donnée ou de chaque classe

b - Ecart-type : L'écart type, noté σx est la racine carré de la moyenne du carré des écarts à la moyenne, c'est à dire la racine carrée de la variance.

$$\delta_x = \sqrt{(\delta_x)^2} = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{N}}$$

L'écart-type est une mesure de dispersion par rapport à la moyenne qui intègre les valeurs algébriques des écarts à la moyenne et qui pourra, à ce titre être réintroduite dans des calculs algébriques ultérieurs. Elle présente de plus l'avantage d'avoir une **signification probabiliste** que ne possède pas l'écart absolu moyen. La théorie des probabilités permet en effet d'estimer la chance qu'à une valeur d'être éloignée de la moyenne de plus d'un certain nombre d'écart-types.

Lorsqu'une distribution est **gaussienne** (on dit aussi "**normale**") les probabilités de trouver les valeurs a une distance donnée de la moyenne sont les suivantes :

68.3 % des valeurs sont comprises entre $(\bar{x}-\sigma_x)$ et $(\bar{x}-\sigma_x)$

95.5 % des valeurs sont comprise entre $(\bar{x}-2\sigma_x)$ et $(\bar{x}+2\sigma_x)$

99.7 % des valeurs sont comprises entre $(\bar{x}-3\sigma_x)$ et $(\bar{x}+3\sigma_x)$

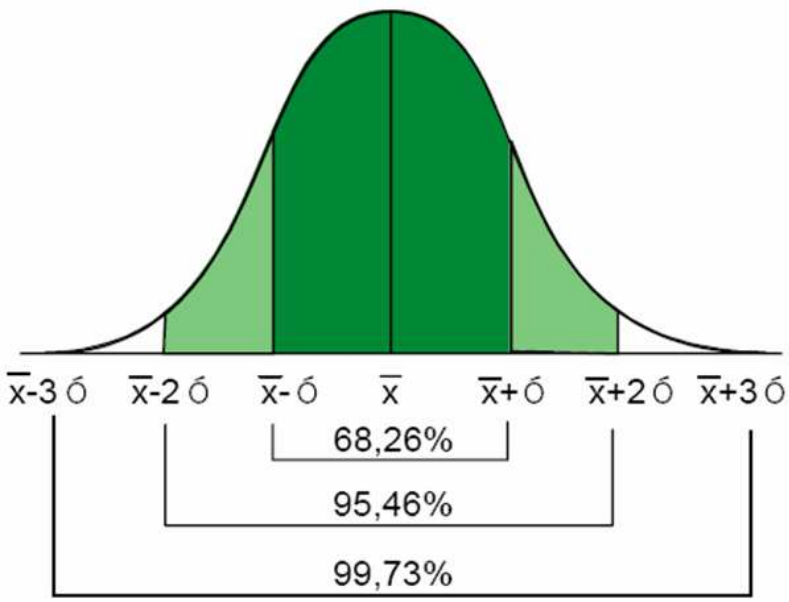


***Figure 5**: Représentation graphique d'une distribution « normale » (Loi de Gauss ou Loi Normale)*

L'écart-type pour des données rangées ou groupées en classe devient :

$$\delta_x = \sqrt{(\delta_x)^2} = \sqrt{\frac{\sum_{i=1}^{n} ni\ (xi - \bar{x})^2}{N}}$$

NB : sur des échantillons de faible taille (n<30), on utilise l'écart-type modifié, soit en divisant par **n-1** au lieu de n (les calculatrices le font automatiquement).

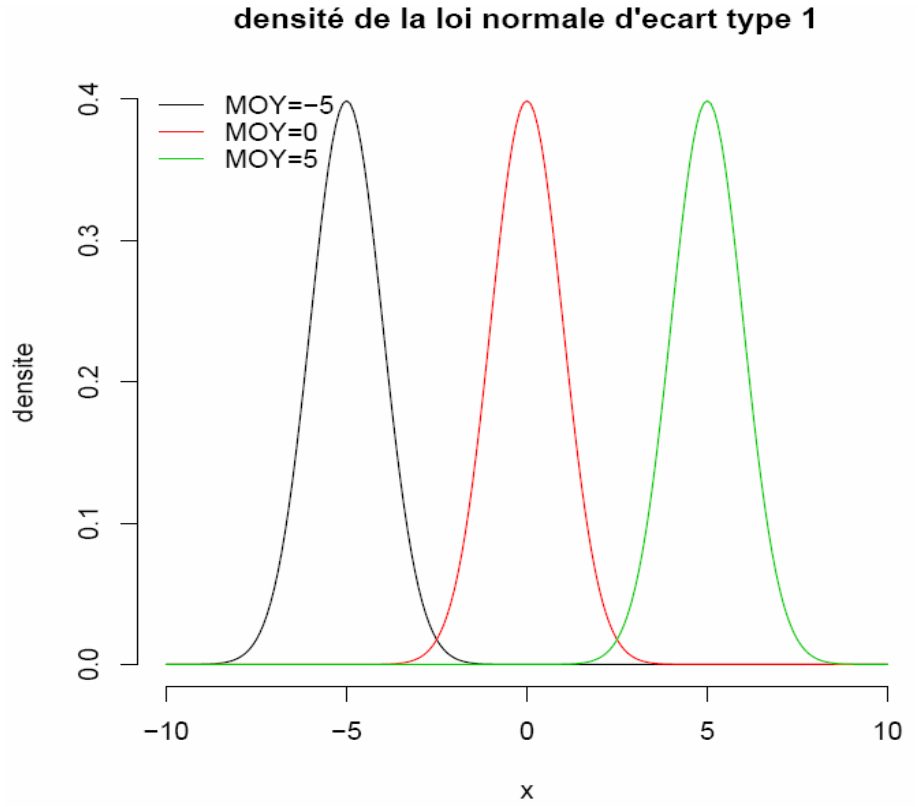


Figure 6: *Densités de lois gaussiennes ayant une même variance mais des moyennes différentes*

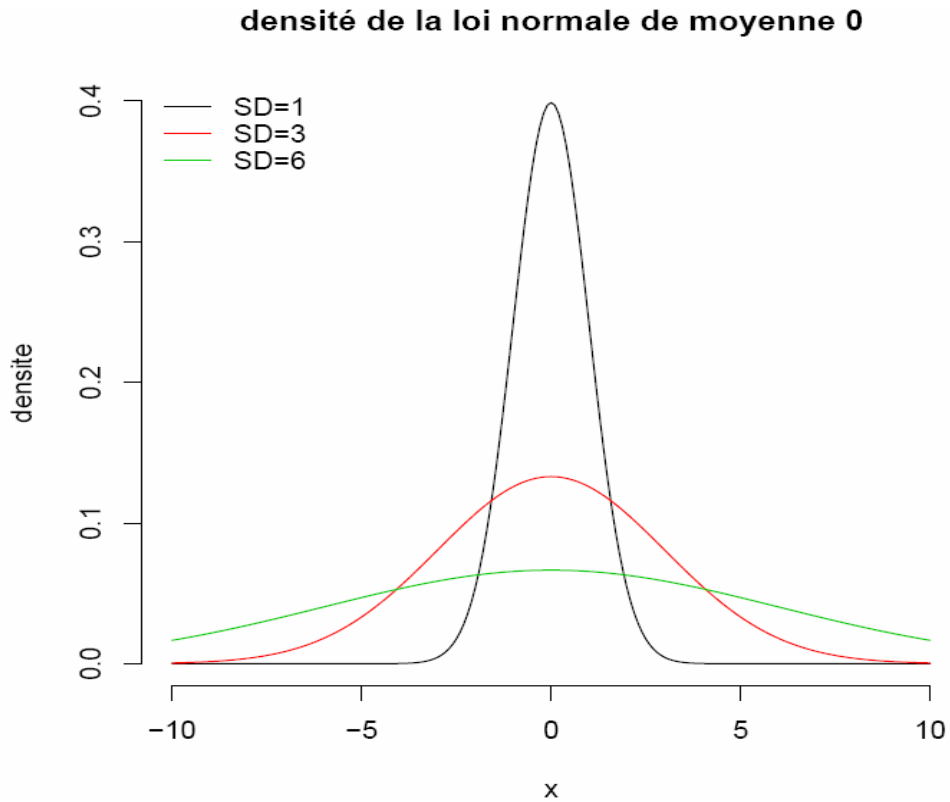


Figure 7 : *Densités de lois gaussiennes ayant une même moyenne mais des variances différentes*

c- Simplification des écritures des variances (respectivement des écart-types)

La formule de la variance peut être remplacée par une formule plus facile à utiliser (**formule pratique de calcul**) à savoir :

$$(\delta_x)^2 = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{N} = (\delta_x)^2 = \frac{\sum_{i=1}^{n} xi^2 - \frac{(\sum_{i=1}^{n} xi^2)}{N}}{N}$$

Démonstration :

Rappel : $(a-b)^2 = a^2 - 2ab + b^2$

de même $(xi-\bar{x})^2 = xi^2 - 2xi\bar{x} + \bar{x}^2$ ainsi $\sum_{i=1}^{n}(xi-\bar{x})^2 = \sum_{i=1}^{n}(xi^2 - 2xi\bar{x} + \bar{x}^2)$

du faite que la moyenne est une constante, la formule peut s'écrire :

$$\sum_{i=1}^{n}(xi-\bar{x})^2 = \sum_{i=1}^{n}(xi^2 - 2xi\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} xi^2 - 2\bar{x}\sum_{i=1}^{n} xi + N\bar{x}^2$$

Or : $\bar{x} = \frac{\sum_{i=1}^{n} xi}{N}$ il suffit alors de le remplacer par sa valeur dans la dernière expression :

$$\sum_{i=1}^{n}(xi-\bar{x})^2 = \sum_{i=1}^{n} xi^2 - 2\bar{x}\sum_{i=1}^{n} xi + N\bar{x}^2 = \sum_{i=1}^{n} xi^2 - 2(\frac{\sum_{i=1}^{n} xi}{N})\sum_{i=1}^{n} xi + N\left(\frac{\sum_{i=1}^{n} xi}{N}\right)^2$$

$$\sum_{i=1}^{n}(xi-\bar{x})^2 = \sum_{i=1}^{n} xi^2 - 2\bar{x}\sum_{i=1}^{n} xi + N\bar{x}^2 = \sum_{i=1}^{n} xi^2 - 2\frac{\left(\sum_{i=1}^{n} xi\right)^2}{N} + \frac{\left(\sum_{i=1}^{n} xi\right)^2}{N}$$

$$\sum_{i=1}^{n}(xi-\bar{x})^2 = \sum_{i=1}^{n} xi^2 - \left(\frac{(\sum_{i=1}^{n} xi)^2}{N}\right) \text{ ainsi } \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{N} = \frac{\sum_{i=1}^{n} xi^2 - \left(\frac{(\sum_{i=1}^{n} xi)^2}{N}\right)}{N}$$

Pour des données condensées <u>la formule pratique de calcul</u> devient :

$$\frac{\sum_{i=1}^{n} ni\, xi^2 - N\left(\frac{(\sum_{i=1}^{n} xi)}{N}\right)^2}{N}$$ ce qui équivaux à $$\frac{\sum_{i=1}^{n} ni\, xi^2}{N} - \left(\frac{(\sum_{i=1}^{n} xi)}{N}\right)^2 = \frac{\sum_{i=1}^{n} ni\, xi^2}{N} - \bar{x}^2$$

Attention – Remarque
Dans le cas d'un échantillon la formule de la variance devient :

$$\frac{\sum_{i=1}^{n} (xi - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} xi^2 - \left(\frac{(\sum_{i=1}^{n} xi)}{n}\right)^2}{n-1}$$

Seule la valeur N du dénominateur commun est remplacée par n-1

6.2.2. Les paramètres de dispersion relative

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur. Dans le cas contraire, la comparaison n'est possible qu'en ayant recours à des mesures de **dispersion relative**, c'est à dire en effectuant le rapport entre un paramètre de dispersion absolue et la valeur centrale qui lui tient de référence.

Un paramètre de dispersion relative est une mesure de **l'écart relatif** des valeurs d'une distribution à une valeur centrale. C'est donc le rapport d'un paramètre de dispersion absolue divisé par une valeur centrale. On obtient un nombre sans dimension qui peut être exprimé en %.

Dispersion relative = Paramètre de dispersion absolue/Valeur centrale

- le coefficient interquartile relatif

= $(X_{q3}-X_{q1})$/ médiane X

- l'écart moyen relatif

= E.A.M. / $\bar{X}$

- le coefficient de variation

= σ_x / $\bar{X}$

Remarque très importante : Le calcul d'un paramètre de dispersion relative n'est possible que pour les caractères quantitatifs **positifs** (toutes les modalités sont des nombres positifs).

Explication des paramètres de dispersion relative pour la variance et l'écart-type :
Ces deux mesures de dispersion (variance et écart-type) sont des grandeurs de même ordre de la variable étudiée : il s'agit d'une mesure de dispersion absolue or pour comparer des séries différentes, il faut éliminer l'unité de mesure afin d'obtenir une mesure de dispersion relative on utilise alors le coefficient de variation exprimé en % :

$$Cv = 100 \frac{\delta}{\bar{x}}$$

Plus le coefficient de variation est faible, plus la dispersion est faible.

6.3 Exercices d'Applications avec explicitation et utilisation du logiciel Excel

APPLICATION VI

Application VI.1 - Exemple n° 1 : Soit un tableau de données issu d'une nécropole :

nombres d'objets	nombre de tombes	nombre total d'objets
1	1	1
2	22	44
3	15	45
4	20	80
5	9	45

1 - Calculez :
a• Le nombre moyen d'objets par tombe
b• Le nombre médian d'objets par tombe
c• L'écart-type du nombre d'objets
d• Le coefficient de variation
2 - Tracez l'histogramme

Réponse :
1- Calculs
a- Le nombre moyen d'objets par tombe, revient à rechercher **la moyenne arithmétique** : elle est égale au nombre total d'objets divisé par le nombre total de tombes, soit X = 215/67 = 3,2 objets par tombe
b- Le nombre médian d'objets par tombe : revient à rechercher **la médiane** dans une série impaire de type (N = 2p + 1), la tombe de rang p + 1, soit la 34ème tombe donne ce nombre médian, soit 3 objets par tombe
M = 3 : la moitié des tombes possèdent de 1 à 3 objets et la moitié de 3 à 5 objets.

Explicitation des calculs :
- Calculons le nombre de tombes = 67 tombes
- Recherchons le rang = (67+1)/2 = 34 ème tombes
- Retrouvons le nombre d'objets dans la 34ème tombe

Pour ce faire il est possible de calculer les effectifs cumulés des tombes

nombres d'objets	nombre de tombes	Effectifs cumulés de tombes
1	1	1
2	22	23
3	15	38
4	20	58
5	9	67

Il y a 3 objets de la 23ème tombe à la 38ème tombe. Donc dans la 3ème tombe il y a 3 objets. La médiane est donc égale à 3 objets par tombe.

c- l'écart-type du nombre d'objets : δ

L'écart type est la racine carrée de la variance, calculons la variance

- La variance (δ^2) est égale à la somme des différences au carré entre le nombre d'objets de chaque tombe et le nombre moyen d'objets divisée par le nombre total de tombes :

δ^2 = ((1-3,2)2+ 22 x ((2-3,2)2)+ 15 x ((3-3,2)2)+ 20 x ((4-3,2)2)+ 9 x (5-3,2)2)) / 67
= 4,84 + 31,68 + 0,6 + 12,8 + 29,16 / 67
= 79,08 / 67 = 1,18

L'écart type est donc $\delta = \sqrt{1,18} = 1,08$ objets

d- Le coefficient de variation exprimé en pourcentage est égal à 100 fois l'écart-type divisé par la moyenne :

Cv = 100 x (1,08/ 3,2) = 33,75 %

Ce coefficient montre une faible dispersion des valeurs autour de la moyenne.

Exemple de tableau obtenu sur Excel

nombre d'objets (xi)	nombre de tombes (pi)	Effectifs cumulés des tombes	Total partiel des objets (pixi)	(xi-X)	$(xi-X)^2$	$pi(xi-X)^2$
1	**1**	1	1	-2,21	4,88	4,88
2	**22**	23	44	-1,21	1,46	32,15
3	**15**	38	45	-0,21	0,04	0,65
4	**20**	58	80	0,79	0,63	12,52
5	**9**	67	45	1,79	3,21	28,87
Total des tombes	**67**	**Total des objets**	**215**		**Somme**	**79,07**
Moyenne (objets / tombe)	**3,2**				**Variance**	**1,18**
Médiane (objets / tombe)	**3**				**Ecart-type (objet)**	**1,09**
					Cœff. de variation	**33,85%**

<u>Application VI.2</u> - Exemple n° 2

Le nombre d'ossements conservés a été enregistré pour dix sites :

site 1 : 2011	**site 6 : 1151**
site 2 : 502	**site 7 : 322**
site 3 : 1107	**site 8 : 903**
site 4 : 355	**site 9 : 2535**
site 5 : 2983	**site 10 : 16631**

Calculez successivement pour les sites 1 à 10 puis 1 à 9
- Le nombre moyen d'ossements par site
- Le site médian de la distribution
- L'écart-type
- Le coefficient de variation
- Tirer des conclusions

Eléments de réponse : Pour calculer la médiane manuellement: il faut ordonner les séries en fonction du nombre d'ossements.

Nombre d'ossements	Séries classées par ordre d'ossements croissant	
322	**S7**	**Rang 1**
355	**S4**	**Rang 2**
502	**S2**	**Rang 3**
903	**S8**	**Rang 4**
1107	**S3**	**Rang 5**
1151	**S6**	**Rang 6**
2011	**S1**	**Rang 7**
2335	**S9**	**Rang 8**
2983	**S5**	**Rang 9**
16631	**S10**	**Rang 10**

- **Pour les 10 séries** (nombre paire de séries) :
Le rang est entre 5 et 6 soit la médiane = (1107+1151)/2 = 1129 ossements

- **Pour les 9 séries** (nombre impaire de séries) :
Le rang est (9+1)/2 soit 5, la médiane correspondante est donc = 1107 ossements

Eléments de correction en utilisant Excel

Séries	xi	Nombre de séries	10
S1	2011	Somme	28500
S2	502	Médiane	1129
S3	1107	Moyenne	2850
S4	355	Variance	24286796,44
S5	2983	Ecartype	4928,16
S6	1151	Coeff. Variation	172,92%
S7	322	**Nombre de séries**	**9**
S8	903	Somme	11869
S9	2535	Médiane	1107
S10	16631	Moyenne	1318,78
		Variance	945429,19
		Ecartype	972,33
		Coeff. Variation	73,73%

APPLICATION VII : Approfondissement

EXEMPLE DE CALCULS PAR CHANGEMENT D'ORIGINE

*** Calcul de la moyenne par changement d'origine et d'unité.**

Dans certains cas la recherche de la moyenne d'une série statistique entraîne des calculs lourds et fastidieux. Dans de telles situations, il est conseillé, voire utile, d'effectuer un changement de variable (ou de code) permettant d'accélérer et de simplifier le calcul. Ce changement peut à la fois toucher l'origine et l'unité. Par exemple il est possible d'effectuer une transformation linéaire de la forme **y = ax + b**

L'objectif est de rechercher la meilleure valeur de x pour simplifier au mieux les calculs.

- La méthode des ζ (lire ksi) **dans le cas de données réparties en classes**

(Par commodités d'écriture ζ peut être remplacé par la lettre u, ou autres lettre (il suffit juste de le préciser))

Dans cette méthode $\mathbf{y = \zeta = ax+b}$ et les paramètres **a** et **b** de la transformation sont choisis de façon à ce que $\zeta = \frac{1}{i}x - \frac{xo}{i}$

Avec :

* $\frac{1}{i}$ est le paramètre **« a »** de la transformation **y = ax + b ;**

*** i** est la largeur de la classe (valeur séparant les valeurs centrales de 2 classes consécutives)

*** -** $\frac{xo}{i}$ est le paramètre **« b »** de la transformation **y = ax + b ;**

***** xo est la valeur centrale de la classe centrale

Cette formule $\zeta = \frac{1}{i}x - \frac{xo}{i}$ **peut aussi être écrite** $\zeta = \frac{x - xo}{i}$ **ou encore** $\zeta = \frac{mx - mxo}{i}$

Avec

- $\mathbf{m_x}$ est la valeur centrale de chaque classe
- $\mathbf{m_{xo}}$ est la valeur centrale de la classe centrale

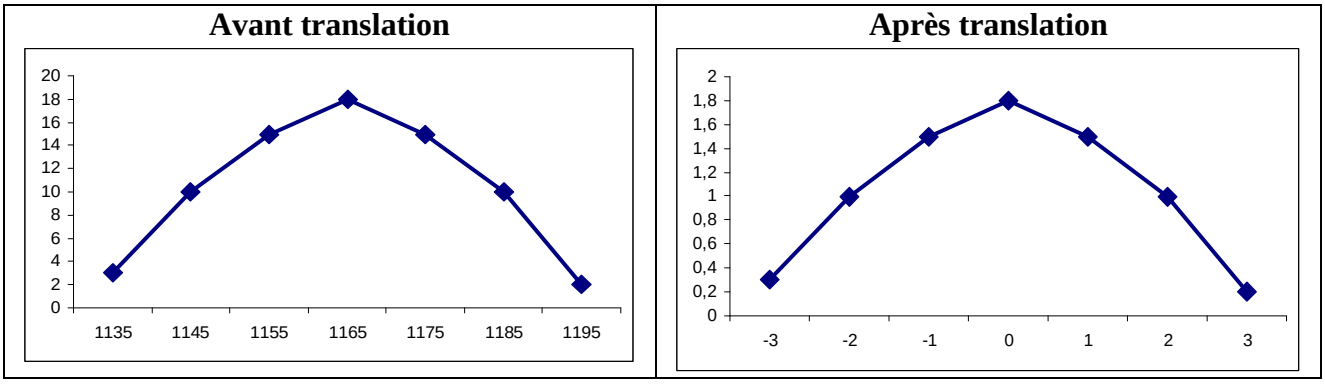
Explicitation et avantage de la transformation selon la méthode des ζ (lire ksi)

En effectuant la transformation $\zeta = \frac{mx - mxo}{i}$ deux opérations sont simultanément réalisées sur la distribution initiale :

Opération 1 : Une translation de la courbe initiale sur l'axe des x qui amènera le sommet approximativement au niveau de l'origine des coordonnées. Ce déplacement est obtenu en retranchant une même valeur xo de chacune des N mesures

Opération 2 : une concentration de la distribution des mesures autour de la moyenne, puisque toutes les différences x-xo sont divisées par i (qui correspond à la largeur de la classe).

Explicitation graphique



Calcul des valeurs de ζ

Avec cette dernière formule $\zeta = \frac{mx - mxo}{i}$, pour chaque x il est alors possible de calculer les différentes valeurs de ζ.

Différents calcul de x

Pour revenir à x il suffira donc de connaître i , xo et ζ de partir de la formule suivante $\zeta = \frac{x - xo}{i}$ et de calculer $x = i\zeta + xo$

- Pour calculer la moyenne $\mu x(\overline{X})$:

La moyenne des ζ se calcule selon la formule $\mu\zeta = \frac{\sum_{ni} \zeta}{N}$

Ce qui est recherché ce n'est pas la $\mu\zeta$ mais la moyenne μx

Il faut donc revenir au système initial :

Si $\zeta = \frac{x - xo}{i}$ il est facile de comprendre que $\mu\zeta = \frac{\mu x - xo}{i}$

C'est-à-dire que la moyenne d'un échantillon d'une population x peut-être calculée selon l'équation suivante : $\mu x = \mu\zeta i + xo$

*** Pour calculer la variance V (δ^2) et l'écart-type δ :**

Selon le même principe Vx est calculé selon la formule suivante :

$(\delta_x)^2 = Vx = i^2 (\delta\zeta)^2 = i^2 V\zeta$ (pour des commodités d'écriture ζ est parfois remplacée par u)

$$V\zeta = (\delta\zeta)^2 = \frac{\sum_{i=1}^{n} (\zeta i - \mu\zeta)^2}{N}$$

Pour les données rangées en classes :

$$V\zeta = (\delta\zeta)^2 = \frac{\sum_{i=1}^{n} ni\ (\zeta i - \mu\zeta)^2}{N}$$

L'écart-type est obtenu de la façon suivante : $\delta x = \sqrt{Vx} = i\sqrt{V\zeta}$

Explicitation avec exemple 1 :
Calculez la moyenne de la série suivante avec une transformation selon x= iu + xo = **10**u + **1165**

Classes	**1130-1140**	**1140-1150**	**1150-1160**	**1160-1170**	**1170-1180**	**1180-1190**	**1190-1200**
Ni (effectifs respectifs)	3	10	15	18	15	10	2

Eléments de correction :
Les calculs sont reportés dans le tableau suivant :

Classes	**1130-1140**	**1140-1150**	**1150-1160**	**1160-1170**	**1170-1180**	**1180-1190**	**1190-1200**
mi (milieu des classes)	1135	1145	1155	**1165**	1175	1185	1195
ni (effectifs respectifs)	3	10	15	18	15	10	2
ni x mi	3405	11450	17325	20970	17625	11850	2390
ζ (ksi)	-3	-2	-1	0	1	2	3
ni x ζ	-9	-20	-15	0	15	20	6
changement des ni	0,3	1	1,5	1,8	1,5	1	0,2

Formules et applications numériques

$\sum_{ni} \zeta$ ni = N	(3+10+15+18+15+10+2)	**73**
i (distance interclasse)	= 1145-1135 = 1155-1145=…=1195-1185	**10**
$xo = mi$	Choix de la classe :[1160-1170[→ 1165	**1165**
Equation y= ax+b	a=1/i=0,1 ; b=-xo/i= 1165/10=116,5 Y= 0,1X- 116,5	**Y= 0,1X- 116,5**
$\zeta = \frac{mx - mxo}{i}$	$\frac{1135-1165}{10}; \frac{1145-1165}{10}; \frac{1155-1165}{10}; \frac{1165-1165}{10}; ....$	**-3 ; -2 ; -1 ; 0 ; 1 ; 2 ; 3**
$\sum$ **ni**ζ	(-9-20-15+0+15+20+6)	**-3**
Moyenne $\mu\zeta = \frac{\sum_{ni} \zeta}{N}$	**-3/73**	**-0,04109589**
Moyenne $\mu x = \mu\zeta i + xo$	=-0,04109589x10 +1165	**1164,589041**

Une correction avec calcul et graphe est aussi fournie sur le fichier Excel des corrections

Exemple 2 :
Calculer la moyenne, la variance et l'écart-type de la série statistique suivante :

classe	ni
1200-1250	15
1250-1300	20
1300-1350	38
1350-1400	25
1400-1450	11

Eléments de solution

$\sum_{ni} \zeta$ ni = N	(15+20+38+25+11)	**109**
i (distance interclasse)	= 1250-1200 =…=1450-1400	**50**
$xo = mi$	Choix de la classe :[1300-1350[→ 1325	**1325**
Equation y= ax+b	a=1/i=1/50=0.02 ; b=-xo/i= 1325/50=26,5 $y = \frac{1}{50}x - \frac{1325}{50}$ = 0,02x - 26,5	$y = 0{,}02x - 26{,}5$
$\zeta = \frac{mx - mxo}{i}$	$\frac{1225-1325}{50}; \frac{1275-1325}{50}; \frac{1325-1325}{50}; \frac{1375-1325}{50}; \frac{1425-1325}{50}$	**-2 ; -1 ; 0 ; 1 ; 2**
$\sum$ **ni**ζ	(-30-20+0+25+22)	**-3**
Moyenne : $\mu\zeta = \frac{\sum_{ni} \zeta}{N}$	**-3/109**	**-0,02752294**
Moyenne $\mu x = \mu\zeta i + xo$	=-0,02752294x50+1325	**1323,62385**
Variance de ζ $V\zeta = (\delta\zeta)^2 = \frac{\sum_{i=1}^{n} ni\,(\zeta i - \mu\zeta)}{N}$	$\frac{15\,(-2-(-0{,}0275)^2}{109} + \frac{20\,(-1-(-0{,}0275)^2}{109} + ...$ =148,917/109	**1,36621497**
$(\delta_x)^2 = Vx = i^2(\delta\zeta)^2 = i^2 V\zeta$	50x50x1,36621497	**3415,53741**
$\delta x = \sqrt{Vx} = i\sqrt{V\zeta}$	$\sqrt{3415{,}53741}$	**58,4425993**

Méthode avec changement de variable : x= iu + xo = **50**u + **1325**

classe	ni	centre des xi	i	u	niu	ui-u	(ui-u)2	ni(ui-u)2
1200-1250	15	1225	**50**	**-2**	-30	-1,972	3,89066	58,3599
1250-1300	20	1275	50	**-1**	-20	-0,972	0,94571	18,9142
1300-1350	38	**1325**	50	**0**	0	0,027	0,00075	0,02878
1350-1400	25	1375	50	**1**	25	1,027	1,05580	26,3950
1400-1450	11	1425	50	**2**	22	2,027	4,11084	45,2193
Somme	**109**				**-3**			**148,917**
Moyenne u	**-0,02752294**							
Moyenne x	**1323,62385**							
Vu	**1,36621497**							
Vx	**3415,53741**							
Ecart-type x	**58,4425993**							

Astuces :

Dans la pratique, lors de l'utilisation de la méthode des ζ il suffit directement :

1/ d'affecter la valeur 0 dans la colonne des ζ à la classe la plus centrale. Ensuite à partir de ce 0 central d'affecter les valeurs -1, -2,-3,…, -n dans valeurs ζ des classes plus faibles et +1, +2,+3,…, +n dans les valeurs ζ des classes plus fortes.

2/ d'effectuer les produits de ζ par les effectifs des classes ;

3/ pour le calcul de la moyenne μζ, de faire le total des de ζ et de les diviser par l'effectif total.

4/ d'utiliser les formules pour le calcul des moyennes, des variances et des écart-types

6.3. Paramètres de forme

Ces paramètres permettent de préciser la forme de la distribution expérimentale. Ils affinent la description de la distribution d'une variable et facilite la comparaison de plusieurs distributions expérimentales. Les paramètres de forme que nous aborderons sont :
(1) le coefficient d'asymétrie il permet de nous renseigner sur la façon régulière ou non dont les observations se répartissent de part et d'autre d'une valeur centrale.
(2) le coefficient d'aplatissement dont l'objet et de faire apparaître si une faible variation de la variable entraîne ou non une forte variation des fréquences relatives.

Remarque
On dit qu'une variable est uni-modale si sa distribution ne présente qu'un maximum, bimodale si elle en présente deux.

6.3.1. Coefficient d'asymétrie et de dérive

Le coefficient d'asymétrie renseigne sur l'asymétrie et éventuellement la dérive par rapport à une valeur centrale choisie. La distribution d'une variable est symétrique si les observations sont également dispersées de part et d'autre d'une valeur centrale. Ainsi, dans le cas de distributions symétriques, moyenne et médiane sont confondues, sinon elles sont distinctes.

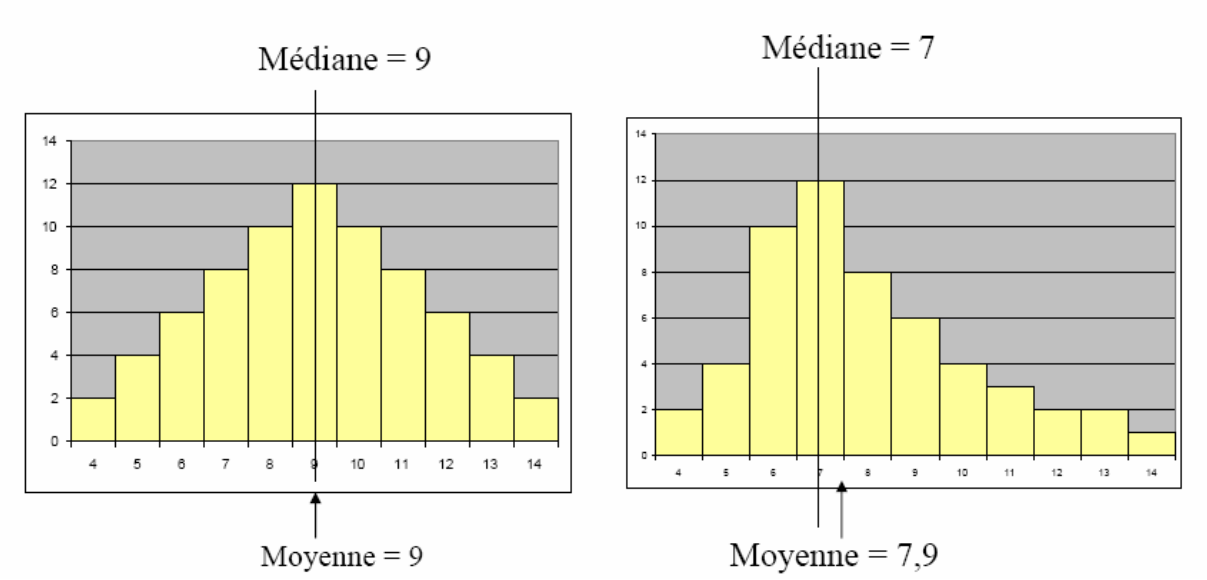


Figure 8 : *Exemple de distributions symétrique et de dissymétrie*

6.3.1.1. Coefficient d'asymétrie

Ce coefficient mesure l'asymétrie d'une distribution, il renseigne sur une asymétrie négative (dissymétrie à gauche), ou une asymétrie positive (dissymétrie à droite), c'est-à-dire il précise si la répartition "penche" d'un côté ou de l'autre. Selon la valeur centrale choisie (mode, médiane ou moyenne arithmétique), il existe différentes manières de caractériser et de mesurer une dissymétrie.

Astuce :
- Dans le cas d'une dissymétrie positive on a généralement (partie droite plus longue que la partie gauche) : **Mo** (Mode) < **Md** (Médiane) < μ (Moyenne)
- Dans le cas d'une dissymétrie négative on a généralement (partie gauche plus longue que la partie droite) : **Mo** (Mode) > **Md** (Médiane) > μ (Moyenne)

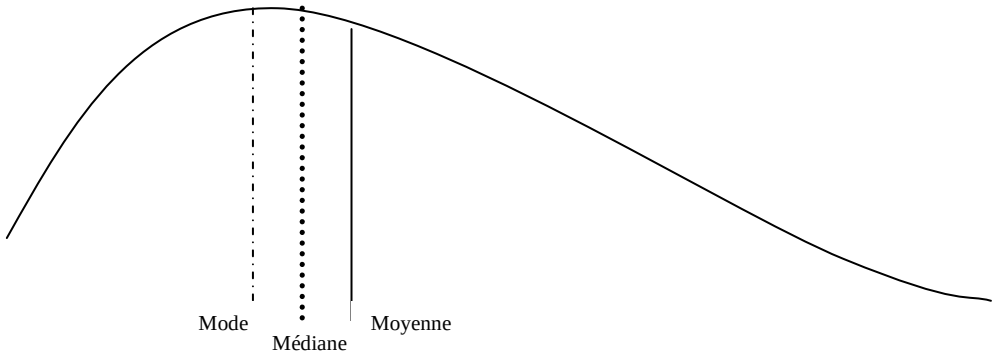


***Figure 9** : Exemple de dissymétrie à droite (distribution étiré à droite et oblique à gauche)*

- Les coefficients d'asymétrie de Yule, si la valeur centrale choisie est la médiane :
Yule propose une mesure de l'asymétrie en comparant l'étalement vers la gauche et l'étalement vers la droite, tous deux repérés par la position des quartiles (Q1, Médiane (Q) et Q3)

$$s = \frac{(Q_3 - Q) - (Q - Q_1)}{(Q_3 - Q) + (Q - Q_1)}$$

Si : S = 0 ⇔ symétrie parfaite
S > 0 ⇔ oblique à gauche (ou étalement à droite)= dissymétrie à droite
S < 0 ⇔ oblique à droite (ou étalement à gauche)= dissymétrie à gauche

- Les coefficients d'asymétrie de Pearson, si les valeurs centrales choisies sont le mode et la moyenne. Pearson propose deux coefficients :

a) le premier coefficient d'asymétrie de Pearson analyse la position de deux valeurs centrales (le mode et la moyenne arithmétique) relativisée par la dispersion de la série :

$$p = \frac{\mu - Mode}{\delta}$$

Si : p = 0 ⇔ symétrie parfaite
p > 0 ⇔ oblique à gauche (ou étalement à droite)= dissymétrie à droite
p < 0 ⇔ oblique à droite (ou étalement à gauche)= dissymétrie à gauche

Remarque : ce coefficient est plutôt performant pour des distributions faiblement asymétriques.

b) le second coefficient d'asymétrie de Pearson (β_1) est plus élaboré : il s'appuie sur le calcul des moments centrés. Il s'écrit :

$$\boxed{\beta_1 = \frac{\mu_3^2}{\mu_2^3}}$$

Où
$\mu_3 = m_3 - 3m_1 m_2 + 2m_1^3$
et $\mu_2 = m_2 - m_1^2 = s^2$

Avec
$m_1 = \frac{\sum n_i x_i}{\sum n_i} = \bar{x}$; $m_2 = \frac{\sum n_i x_i^2}{\sum n_i}$ *et* $m_3 = \frac{\sum n_i x_i^3}{\sum n_i}$

De façon plus générale, on a :

Moment d'ordre r : $m_r = \frac{1}{n}\sum_{i=1}^{k} nix_i^r$

Moment centré d'ordre r : $\mu_r = \frac{1}{n}\sum_{i=1}^{k} ni(x_i - \bar{x})^r$

Si :

$\beta_1 = 0 \Leftrightarrow$ *symétrie*

$\beta_1 > 0 \Leftrightarrow$ *oblique à gauche (ou étalement à droite)= dissymétrie à droite*

$\beta_1 < 0 \Leftrightarrow$ *oblique à droite (ou étalement à gauche)= dissymétrie à gauche*

- Les coefficients d'asymétrie de Fisher, si la valeur centrale choisie est la moyenne :

Fisher propose un coefficient qui correspond à la racine carrée du coefficient β_1 de Pearson :

$$\gamma_1 = \frac{\mu_3}{\delta^3}$$

Où

$s^3 = \sqrt{u_2^3}$

Si :

$\gamma_1 = 0 \Leftrightarrow$ *symétrie*

$\gamma_1 > 0 \Leftrightarrow$ *oblique à gauche (ou étalement à droite)= dissymétrie à droite*

$\gamma_1 < 0 \Leftrightarrow$ *oblique à droite (ou étalement à gauche)= dissymétrie à gauche*

6.3.1.2. Coefficient de dérive

Le coefficient **d'asymétrie de Fisher** calculé ci-dessus correspond pour certains auteurs au **coefficient de dérive** « d » ainsi

$$d = \gamma_1 = \frac{\mu_3}{\delta^3}$$

Les coefficients d et δ sont très sensibles aux fluctuations d'échantillonnage, il faudra disposer d'un grand nombre d'observations pour les utiliser.

APPLICATION VIII

Application VIII.1

Reproduire sur Excel ce tableau de distribution et calculer les coefficients de forme

Classes	n_i	x_i	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$
$50-60$	8	55	440	24 200	1 331 000
$60-70$	10	65	650	42 250	2 746 250
$70-80$	16	75	1 200	90 000	6 750 000
$80-90$	14	85	1 190	101 150	8 597 750
$90-100$	10	95	950	90 250	8 573 750
$100-110$	5	105	525	55 125	5 788 125
$110-120$	2	115	230	26 450	3 041 750
Total	65		5 185	429 425	36 828 625

On trouve :

$$mod \simeq 75$$
$$med \simeq 79.1$$
$$Q_1 \simeq 68.2$$
$$Q_3 \simeq 90.7$$
$$m_1 = \bar{x} = \frac{\sum n_i x_i}{n} = \frac{5\ 185}{65} = 79,8$$
$$m_2 = \frac{\sum n_i x_i^2}{n} = \frac{429\ 425}{65} = 6\ 606,5$$
$$m_3 = \frac{\sum n_i x_i^3}{n} = \frac{36\ 828\ 625}{65} = 566\ 594,2$$
$$\mu_2 = m_2 - m_1^2 = 238,46 \Rightarrow s = 15,44$$
$$\mu_3 = m_3 - 3m_1 m_2 + 2m_1^3 = 1\ 337,31$$

d'où :

$$s = \frac{(Q_3 - med) - (med - Q_1)}{(Q_3 - med) + (med - Q_1)} = \frac{(90,7 - 79,1) - (79,1 - 68,2)}{(90,7 - 79,1) + (79,1 - 68,2)} = 0,03$$

$$p = \frac{\bar{x} - mod}{s} = \frac{4,8}{15,44} = 0,3$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{1\ 788\ 398}{13\ 559\ 592} = 0,131$$

$$\gamma_1 = \frac{\mu_3}{s^3} = \frac{1\ 337,31}{3\ 680,8} = 0,363.$$

$\Rightarrow$ la distribution est donc légèrement oblique à gauche.

6.3.2. Coefficient d'aplatissement

Le coefficient d'aplatissement, par référence à la courbe de la loi normale, indique si la distribution de la variable est leptocurtique (pointue), mésocurtique (normale) ou Ainsi, une distribution est dite aplatie si une forte variation de la variable entraîne une faible variation de la fréquence relative (et inversement).

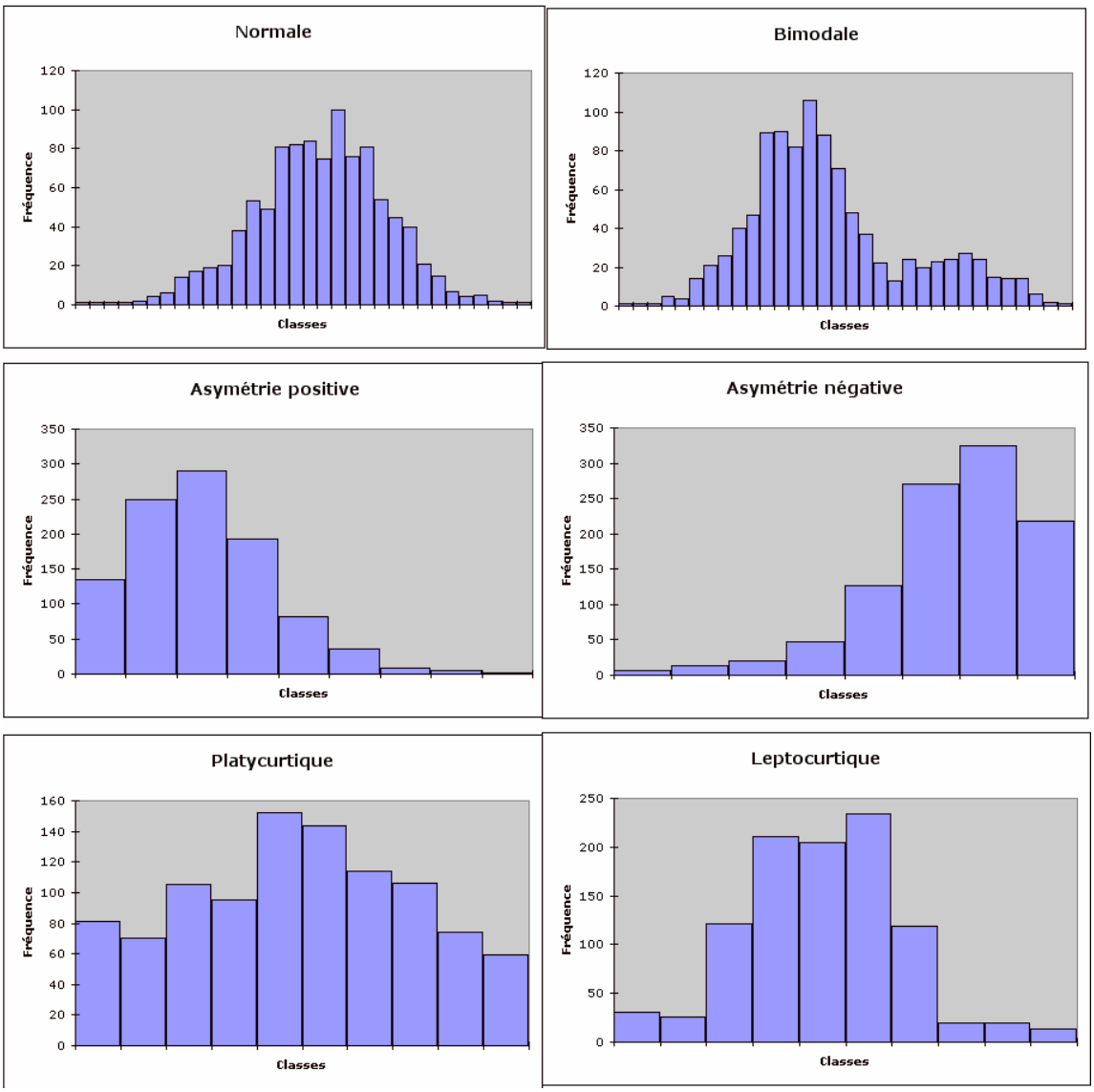


Figure 10 : *Histogrammes illustrant les caractéristiques importantes d'une distribution*

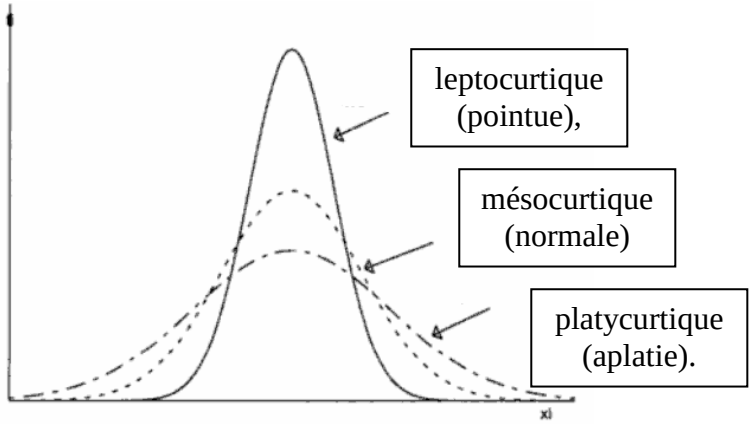


Figure 11 : *Courbe avec coefficient d'aplatissement différent*

- Coefficient d'aplatissement de Pearson

Il s'écrit :

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\delta^4}$$

Ce coefficient est toujours supérieur ou égal à 1. Plus ce coefficient est faible plus la répartition est aplatie (plus la courbe est platicurtique). Plus il est grand, plus les observations sont plus regroupées autour de la moyenne.

β_2 prend la valeur 3 pour une distribution normale.

- Coefficient d'aplatissement de Fisher

Il s'écrit :

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{\delta^4} - 3$$

Si :

$\gamma_2 = 0 \Leftrightarrow$ *distribution normale, l'aplatissement est le m^me que celui de la loi de Gauss réduite*

$\gamma_1 < 0 \Leftrightarrow$ *la distribution est plus aplatie (*platicurtique*)*

$\gamma_1 > 0 \Leftrightarrow$ *la distribution est moins aplatie (*leptocurtique*)*

Application VIII.2

Utiliser les tableaux de distribution suivant et réaliser les calculs sur Excel

Tableau 1 : Distribution

x_i	0	1	2	3
f_i	0,216	0,432	0,288	0,064

On obtient le tableau suivant :

Tableau 2 : Calculs nécessaires (coefficients d'applatissement)

x_i	f_i	$f_i x_i$	$f_i x_i^2$	$(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^3$	$f_i(x_i - \bar{x})^4$
0	0,216	0	0	−1.2	0,311	−0,373	0,448
1	0,432	0,432	0,432	−0,2	0,017	−0,0035	0,00069
2	0,288	0,576	1,152	+0,8	0,184	+0,147	0,11796
3	0,064	0,192	0,576	+1,8	0,207	+0,373	0,6718
$\sum$	1	1,2	2,16		0,72	0,144	1,238
		$m_1 = \bar{x}$	m_2		μ_2	μ_3	μ_4

où

$$\left.\begin{aligned} \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{0,144^2}{0,72^3} = 0,05 \\ \gamma_1 &= \frac{\mu_3}{s^3} = \frac{0,144}{\sqrt{0,72^3}} = 0,24 \end{aligned}\right\} \Rightarrow \text{la distribution est oblique à gauche}$$

$$\left.\begin{aligned} \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{1,238}{0,72^2} = 2,39\ (<3) \\ \gamma_2 &= \frac{\mu_4}{\mu_2^2} - 3 = \frac{1,238}{0,72^2} - 3 = -0,61 \end{aligned}\right\} \Rightarrow \text{la distribution est platicurtique.}$$

PARTIE EXERCICES

Exercices d'applications I : Moyenne, Médiane, Etendue, Quantiles

Etudier les séries suivantes : Série à étudier

Série A :

(rang)	1	2	3	4	5
valeur :	**24.3**	**31.85**	**33.61**	**36.81**	**38.92**

Série B :

(rang)	1	2	3	4	5	6	7	8	9	10
valeur :	**29.2**	**31.4**	**32**	**32.3**	**32.5**	**34.7**	**34.9**	**36.6**	**37.2**	**39.4**

Série C :

(rang)	1	2	3	4	5	6	7	8	9	10	11
valeur :	**28**	**29.06**	**29.09**	**34.49**	**34.92**	**35.76**	**36.73**	**37.21**	**37.28**	**37.68**	**41.17**

Série D :

(rang)	1	2	3	4	5	6	7	8	9
valeur :	**27**	**29**	**30**	**30**	**36**	**36**	**37**	**41**	**42**

Questions : Pour chacune des séries déterminer :

1- Quel est l'effectif de cette série ?
2- Quelle est la médiane de cette série ?
3- Quelle est l'étendue de cette série ?
4- Quelle est la moyenne de cette série ?
5- Quel est le 1er quartile de cette série ?
6- Quel est le 3ème quartile de cette série ?

Réponses

Série A

1- L'**effectif** (***n***) de cet échantillon caractérise sa *taille*; ici, ***n*** = **5**

La **médiane** est un indicateur de position centrale.
Elle correspond à la valeur "centrale" de la série considérée Médiane = 33.61

L'**étendue** est un indicateur de *dispersion* que l'on détermine ainsi :
Etendue = Maximum - Minimum
Ce qui donne ici : Etendue = 38.92 - 24.3 = 14,62

La **moyenne** est une caractéristique de *position "centrale"* qui est déterminée ainsi :

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

L'application de cette formule nous donne ainsi :

n	effectif	= 5
$\sum x_i$	somme des données	= 165.49

$\bar{x}$ moyenne de l'échantillon = 33,1

Les **quartiles** sont aussi des indicateurs de position, ils divisent chacune des partitions définies par la médiane en sous-partitions d'effectifs égaux.

1er quartile Q1 = 31,85

3ème quartile Q3 =36,81

Réponse Série B

Bilan	Résultats attendus
Effectif	10
Médiane	33.6
Etendue	10.2
Moyenne	34
Q_1 = 1er quartile	32
Q_3 = 3ème quartile	36.6

Réponse Série C

Bilan	Résultats attendus
Effectif	11
Médiane	35.76
Etendue	13.17
Moyenne	34.67
Q_1 = 1er quartile	29.09
Q_3 = 3ème quartile	37.28

Réponse Série D

Bilan	Résultats attendus
Effectif	9
Médiane	36
Etendue	15
Moyenne	34
Q_1 = 1er quartile	29.5
Q_3 = 3ème quartile	39

Exercices d'applications II

- REPRESENTATION DE SERIES ET CALCUL STATISTIQUE

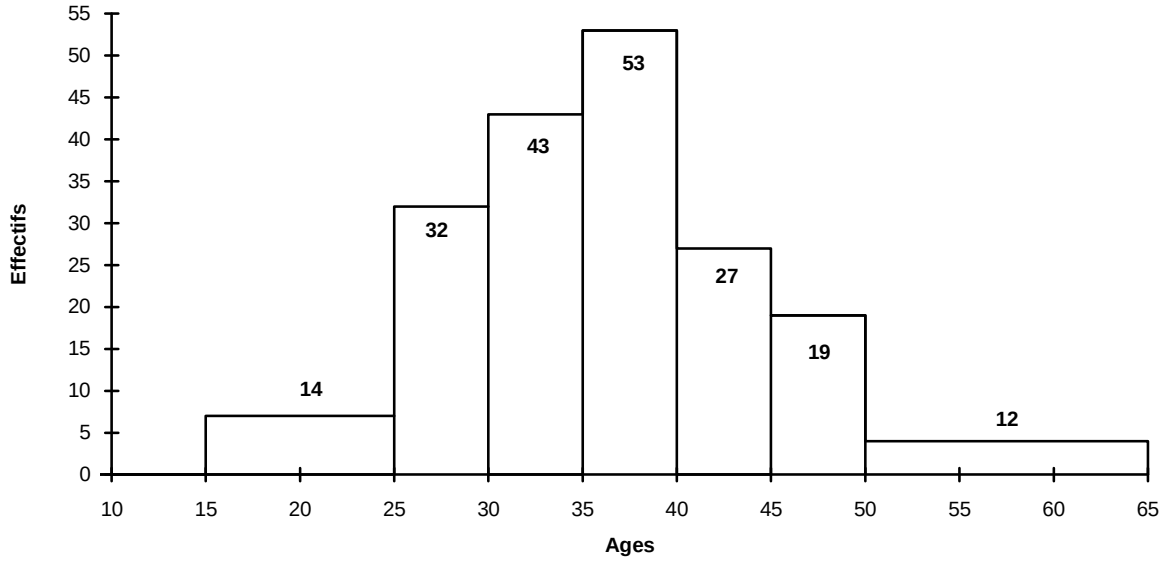
Activités et Applications

I. Tracer l'histogramme représentant une série statistique :

- **Exemple :** répartition suivant leur âge de donneurs de sang

[15,25[	[25,30[	[30,35[	[35,40[	[40,45[	[45,50[
14	32	43	53	27	19

- **Méthode :** on construit des rectangles dont les aires sont proportionnelles aux effectifs des classes correspondantes. La première classe ayant une amplitude double de celle des autres sera représentée par un rectangle de hauteur 2 fois plus petite. De même la dernière classe est représentée par un rectangle de hauteur 3 fois plus petite.
- **Solution :**



II. Établir le tableau permettant d'obtenir les caractéristiques de la série :

- **Solution :**

Ages	Effectifs	Centres	Produits	Cumuls	Cumuls	Carrés	Produits
	ni	xi	ni xi	crois.	décr.	xi2	ni xi2
[15,25[	14	20,0	280,00	14	200	400,00	5 600,00
[25,30[	32	27,5	880,00	46	186	756,25	24 200,00
[30,35[	43	32,5	1 397,50	89	154	1 056,25	45 418,75
[35,40[	53	37,5	1 987,50	142	111	1 406,25	74 531,25
[40,45[	27	42,5	1 147,50	169	58	1 806,25	48 768,75
[45,50[	19	47,5	902,50	188	31	2 256,25	42 868,75
[50,65[	12	57,5	690,00	200	12	3 306,25	39 675,00
	200		7 285,00				281 062,50

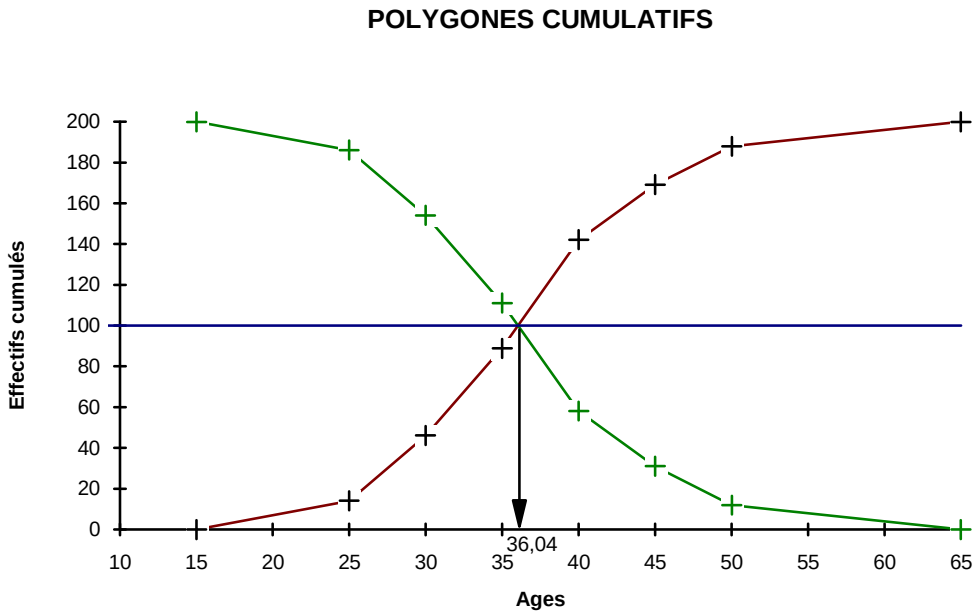
III. Déterminer la moyenne pondérée de la série :

- **Solution :**

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{7285}{200} = 36{,}43 \text{ ans}$$

IV. Déterminer la médiane à l'aide des polygones des effectifs cumulés :

- **Solution :**



V. Calculer la médiane de la série :

- **Solution :**
- demi-effectif : $\frac{200}{2} = 100$
- classe de la 100^{e} personne : $[35;40[$
- rang dans cette classe : $100 - 89 = 11$
- amplitude de cette classe : $40 - 35 = 5$
- effectif de cette classe : 53
- médiane : $Q_2 = 35 + \frac{5 \times 11}{53} = 36{,}04$

VI. Calculer l'écart-type de la série :

- **Méthode :** on utilise l'une des formules suivantes

$$\sigma = \sqrt{\frac{\sum n_i \left(x_i - \bar{x}\right)^2}{\sum n_i}} = \sqrt{\frac{\sum n_i x_i^2}{\sum n_i} - \bar{x}^2}$$

- **Solution :**

Utilisons ici la deuxième formule :

$$\sigma = \sqrt{\frac{281062{,}50}{200} - (36{,}43)^2}$$

$$\sigma = \sqrt{1405{,}3125 - 1327{,}1449}$$

$$\sigma = \sqrt{78{,}1676} = 8{,}841$$

CHAPITRE II

ETUDE DE DEUX VARIABLES STATISTIQUES
- SERIE STATISTIQUE DOUBLE -

1. PRESENTATION D'UNE SERIE A DEUX VARIABLES

L'objectif de cette étude statistique est d'étudier sur une même population de N individus, deux caractères différents (ou modalités différentes) et de rechercher s'il existe un lien ou corrélation entre ces deux variables.

Exemple de relations possibles entre les variables suivantes : taille et âge ; diabète et poids ; taux de cholestérol et régime alimentaire ; niche écologique et population ; ensoleillement et croissance végétale ; toxine et réaction métabolique ; survie et pollution ; effets et doses; organe 1 et 2 ; organe et fonction biologique ; …

Les caractères étudiés peuvent être aussi bien qualitatifs que quantitatifs.

Les résultats sont généralement représentés sous forme d'un **tableau à double entrée**, appelé **tableau à deux dimensions**, ou **tableau croisé** ou **tableau de contingence**, ou parfois **tableau de corrélation**.

Exemple de tableau de contingence

	Effets de doses (variable y)			
Sexe (variable x)	**Effet 1**	**Effet 2**	**Effet 3**	**total**
H	43	36	3	**Total des H** : 82
F	49	12	12	**Total des F** : 73
Total	**Total effet 1 :** 92	**Total effet 2 :** 48	**Total effet 3 :** 15	**Total des H et F** : 155

Effets de doses selon le sexe H ou F

2. GENERALISATION DES REPRESENTATIONS

Désignons par (X, Y) le couple de caractères étudiés.
A chaque observation conjointe (x_i, y_j) est associée le nombre d'individus ayant simultanément la valeur xi pour le caractère X et la valeur yj pour le caractère Y. Ce nombre est noté n_{ij} et appelé **l'effectif associé à l'observation (x_i,y_j)**.

y_j / X_i	y_1	y_2	..	y_j	..	y_z	TOTAL	
x_1	n_{11}	n_{12}		n_{1j}		n_{1z}	$\sum_{j=1}^{z} n1j = n1.$	EFFECTIFS MARGINAUX DE X
x_2	n_{21}	n_{22}		n_{2j}		n_{2z}	$\sum_{j=1}^{z} n2j = n2.$	
….								
x_i	n_{i1}	n_{i2}		n_{ij}		n_{iz}	$\sum_{j=1}^{z} nij = ni.$	
x_i								
….								
x_k	n_{k1}	n_{k2}		n_{kj}		n_{kz}	$\sum_{j=1}^{z} nkj = nk.$	
Total	$\sum_{i=1}^{k} ni1 = n.1$	$\sum_{i=1}^{k} ni2 = n.2$		$\sum_{i=1}^{k} nij = n.j$		$\sum_{i=1}^{k} niz = n.z$	N	
	Effectifs Marginaux de Y							

La ligne et la colonne total correspondent aux marges du tableau.

3. CALCUL DES FREQUENCES D'UNE STATISTIQUE A DEUX VARIABLES

3.1. Fréquences relatives partielles

La fréquence de l'observation (x_i, y_j) s'exprime par l'expression fij. Elle correspond à la proportion d'individus qui possèdent simultanément les valeurs x_i et y_j. Elle est obtenue par la formule suivante :

$$f_{ij} = \frac{n_{ij}}{N} \quad \text{il est a remarquer que} \quad \sum_{i=1}^{k}\sum_{j=1}^{z} f_{ij} = 1$$

3.2. Fréquences relatives marginales fi . et f.j

Il s'agit des fréquences relatives des distributions marginales.

$$f_{i.} = \frac{n_{i.}}{N} \quad et \quad f_{.j} = \frac{n_{.j}}{N}$$

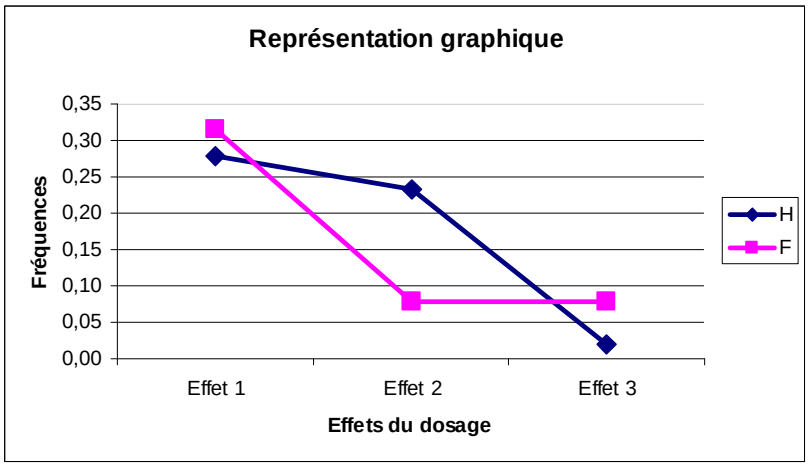
Exemple

	Effets de doses (variable y)			
Sexe (variable x)	**Effet 1**	**Effet 2**	**Effet 3**	**total**
H	43	36	3	**Total des H** : 82
F	49	12	12	**Total des F** : 73
Total	**Total effet 1 :** 92	**Total effet 2 :** 48	**Total effet 3 :** 15	**Total des H et F** : 155

Tableau des fréquences de l'effet des doses selon le sexe

Sexe	Effet 1	Effet 2	Effet 3	total
H	(43/155) = **0,28**	(36/155) = 0,23	(3/155) = 0,02	(82/155) = **0,53**
F	(49/155) = 0,32	(12/155) = 0,08	(12/155) = 0,08	(73/155) = **0,47**
Total	(92/155) = **0,59**	(48/155) = **0,31**	(15/155) = **0,10**	**1**

Représentation graphique :



4. CALCUL DES MOYENNES MARGINALES D'UNE STATISTIQUE A DEUX VARIABLES

Dans certaines distributions statistiques bidimensionnelles il est possible de calculer les moyennes, les variances et les écart-types marginaux. Nous expliciterons ces calculs à travers un exemple.
Soit la série statistique bidimensionnelle du couple (X, Y) suivante :

	Y			
X	**-2**	**0**	**2**	**3**
2	3	4	0	6
3	4	3	3	2
4	2	3	3	2

Calculer respectivement:
1- Les moyennes marginales de X puis de Y
2- Les variances et l'écart-type marginaux de X puis de
3- La moyenne conditionnelle de X quand Y=2
4- La moyenne conditionnelle de Y quand X=3

1 et 2 - Les moyennes marginales, c'est le calcul des moyennes des effectifs marginaux.
- les variances et les écart-types marginaux se calculs aussi sur les effectifs marginaux.
Les formules respectives seront utilisées :

Pour les moyennes : a) $\overline{x} = \dfrac{\sum_{i=1}^{n} xi}{N}$ et b) $\overline{y} = \dfrac{\sum_{j=1}^{n} yj}{N}$

Pour les variances : c) $\boxed{\delta_x^{\,2} = \dfrac{\sum_{i=1}^{n} ni\, xi^2}{N} - \overline{x}^2}$ et d) $\boxed{\delta_y^{\,2} = \dfrac{\sum_{j=1}^{n} ni\, yj^2}{N} - \overline{y}^2}$

Dressons le tableau des distributions afin de faciliter les calculs

	Y				Pour les paramètres de la variable X		
X	**-2**	**0**	**2**	**3**	**ni.**	**xi2**	**ni.xi2**
2	3	4	0	6	**13**	4	52
3	4	3	3	2	**12**	9	108
4	2	3	3	2	**10**	16	160
n.j	**9**	**10**	**6**	**10**	**35**	**somme**	**320**
	Pour les paramètres de la variable Y						
yj2	4	0	4	9	**somme**		
n.jyj2	36	0	24	90	**150**		
Moy marginale X	**2,91**	**Var X**	0,64	**écart-type X**	0,80		
Moy marginale Y	**0,68**	**Var Y**	3,81	**écart-type Y**	1,95		

Applications numériques

a) $\bar{x} = \dfrac{\sum_{i=1}^{n} xi}{N} = \dfrac{(2x13)+(3x12)+(4x10)}{35} = \dfrac{102}{35} = 2{,}91$

b) $\bar{y} = \dfrac{\sum_{j=1}^{n} yj}{N} = \dfrac{(-2x9)+(0x10)+(2x6)+(3x10)}{35} = \dfrac{24}{35} = 0{,}68$

c) $\delta_x^{\,2} = \dfrac{\sum_{i=1}^{n} ni\, xi^2}{N} - \bar{x}^2 = \dfrac{(13x4+12x9+10x16)}{35} - (2{,}91)^2 = \dfrac{320}{35} - 8{,}46 = 0{,}64$

d) $\delta_y^{\,2} = \dfrac{\sum_{i=1}^{n} nj\, yj^2}{N} - \bar{y}^2 = \dfrac{(150)}{35} - (0{,}68)^2 = 3{,}81$

(Correction détaillée sur fichier Excel)

3 – Pour déterminer la moyenne conditionnelle de X quand Y=2, il suffit d'observer le comportement de X relatif à la colonne Y=2

X	**Y= 2**
2	0
3	3
4	3
n.j	**6**

$$\bar{x}_{y=2} = \frac{\sum_{i=1}^{n} xi}{N} = \frac{(0x2)+(3x3)+(3x4)}{10} = \frac{21}{10} = 2{,}1$$

4 – Pour déterminer la moyenne conditionnelle de Y quand X=3, il suffit d'observer le comportement de Y relatif à la colonne X=3

	Y				
	-2	**0**	**2**	**3**	**ni.**
X=3	4	3	3	2	**12**

$$\bar{y}_{x=3} = \frac{\sum_{j=1}^{n} yj}{N} = \frac{(-2x4)+(0x3)+(2x3)+(3x2)}{12} = \frac{4}{12} = 0{,}33$$

5. COVARIANCE

Une première approche entre de la relation éventuelle des valeurs d'une variable X avec des valeurs d'une variable Y est donnée par le calcul de la covariance. La covariance du couple (X, Y), notée Cov (X,Y) correspond à la moyenne de $(X-\overline{X})(Y-\overline{Y})$

La formule est donc la suivante :

$$Cov = \frac{\sum_{i=1}^{n}(xi-\overline{x})(yi-\overline{y})}{N}$$

Analogue à la combinaison des deux formules suivantes

$$(\delta_x)^2 = \frac{\sum_{i=1}^{n}(xi-\overline{x})^2}{N} = \frac{\sum_{i=1}^{n}(xi-\overline{x})(xi-\overline{x})}{N} \qquad (\delta_y)^2 = \frac{\sum_{i=1}^{n}(yi-\overline{y})^2}{N} = \frac{\sum_{i=1}^{n}(yi-\overline{y})(yi-\overline{y})}{N}$$

Dans cette formule la « co-variance » apparaît bien comme une combinaison de la variance de X et celle de Y.

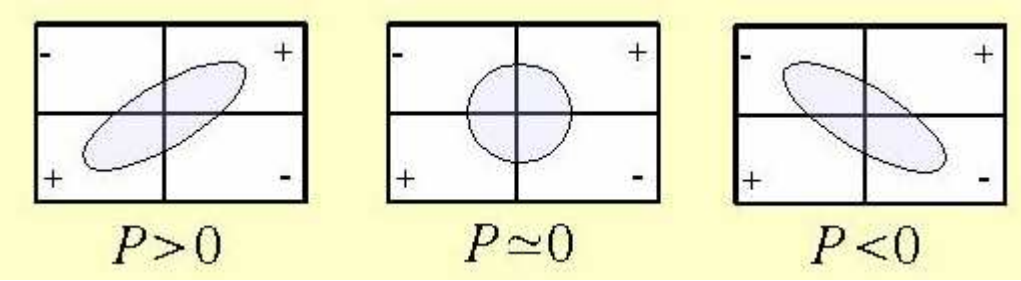
Par analogie aux formules précédentes les formules pratiques de calculs de la covariance peuvent aussi s'écrire :

- Pour des données non groupées : $Cov = \frac{\sum_{i=1}^{n} xiyi}{N} - \overline{x}\,\overline{y}$

- Pour des données groupées : $Cov = \frac{\sum_{i=1}^{n} nixiyi}{N} - \overline{x}\,\overline{y}$

Propriétés de la covariance

- Cov (X, X) = var (X)
- $|Cov(X,Y)| \leq \delta(X)\delta(Y)$
- Le signe de la Cov est un indicateur de la tendance de la relation sens positif ou négatif (direction d'étirement du nuage de point)
-



Une covariance positive indique une tendance « croissante » des valeurs de Y en fonction de X, une covariance négative une tendance « décroissante »

Exemples de calcul de la covariance :

Exemple 1 : Distribution bimodale dans un tableau de contingence

	Y		
X	**-2**	**0**	**2**
0	4	10	5
2	5	12	4
4	2	7	1

Recherchons la covariance (X,Y)

- dressons le tableau de contingence avec les variables calculées

	Y				
X	**-2**	**0**	**2**	**ni**	**nixi**
0	4	10	5	19	0
2	5	12	4	21	42
4	2	7	1	10	40
nj	11	29	10	**50**	**82**
njyj	-22	0	20		
moy mar X	1,64				
moy mar Y	-0,04				
Cov (X,Y)	-0,24				

- la moyenne marginale de x

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{N} = \frac{(0x19)+(2x21)+(4x10)}{50} = \frac{82}{50} = 1{,}64$$

- la moyenne marginale de y

$$\bar{y} = \frac{\sum_{j=1}^{n} yj}{N} = \frac{(-2x11)+(0x29)+(2x10)}{50} = \frac{-2}{50} = -0{,}4$$

Cov= 1/50 ((-2x4x0)+(0x10x0)+(2x5x0)+(-2x5x2)+(0x12x2)+(2x4x2)+(-2x2x4)+(0x7x4)+(2x1x4)
Cov (X,Y) = -0,24

Exemple 2 : Distribution bimodale dans un tableau simple.

6. COEFFICIENT DE CORRELATION

La covariance n'est pas un indicateur indépendant de l'ordre de grandeur des variables impliquées (de l'unité employée, par exemple). Le coefficient de corrélation, noté r, permet de résoudre cette difficulté. Ce coefficient pour le couple (X,Y) s'écrit selon la formule suivante :

$$r = \frac{Cov(X,Y)}{\delta x \delta y}$$

où **δx** et **δx** désignent respectivement **l'écart-type** de la série statistique X et celui de la séries statistique Y

Propriété de r :

- **r est toujours compris entre -1 et 1**, c'est une covariance « réduite »
- **quand (|r |= 1),** les points représentatifs des couples (xi, yi), sont parfaitement alignés sur le graphique :
- **quand (|r |est voisin de 1),** il existe une forte corrélation entre X et Y. Néanmoins (attention), ceci ne veut pas dire qu'il existe une relation de cause à effet entre elles.
- **pour r = 1**, la droite de la pente est croissante
- **Si 0 < r < 1, la corrélation est positive**, X et Y varient dans le même sens.
- **Si -1 < r < 0, la corrélation est négative**, X et Y varient dans le sens contraire.
- **pour r = -1, la droite de la pente est décroissante**
- **quand (r = 0),** aucune tendance ne peut être déterminée

7. DROITE DE REGRESSION OU D'AJUSTEMENT

7.1. Importance de l'étude de corrélation entre 2 variables statistiques

L'une des méthodes simple d'étude de corrélation entre 2 variables consiste à rechercher une courbe d'équation y = f(x) qui passe au plus proche de tous les points expérimentaux. Une telle courbe permet d'avoir une idée sur la tendance de la relation entre les variables étudiées et de formuler d'éventuelles prévisions.

7.2. Droite de régression linéaire

Une droite de régression linéaire s'écrit selon l'équation : « **y = ax +b** ». Cette approche de corrélation repose sur l'hypothèse que la relation entre deux variables est de nature linéaire. En faite, il est possible de soupçonner une relation différente entre ces variables :

- courbe de puissance
- courbe exponentielle
- courbe logarithmique,
- courbe hyperbolique, etc…

Cependant, il existe de nombreuses méthodes permettant de « linéariser » un grand nombre de ces courbes. Ainsi, on se retrouve souvent dans des situations où il est alors possible de tester l'existence d'une relation linéaire entre les variables auxiliaires.

En partant de l'équation $y = ax + b$, a et b doivent être choisis convenablement de sorte que la droite passe au plus proche (ou par le plus possible) des points expérimentaux. Pour ce faire, on utilise la méthode des moindres carrés : On cherche les coefficients a et b de la droite qui minimise la somme des carrés des distances entre les points expérimentaux et la droite de régression (les points théoriques).

- le coefficient **a** (pente) se détermine comme suit :

$$\hat{a} = \frac{Cov(X,Y)}{\delta_x^{\,2}}$$

- le coefficient **b** (ordonnée à l'origine) se détermine comme suit :

$$\hat{b} = \bar{y} - \hat{a}\,\bar{x}$$

Ainsi la droite de régression de Y en X a pour équation :

$$y = ax + b = \frac{Cov(X,Y)}{\delta_x^{\,2}}(x - \overline{X}) + \overline{Y}$$

Pour exprimer X en fonction de Y, il suffit d'inverser les rôles de X et Y

$$x = a'y + b' = \frac{Cov(X,Y)}{\delta_y^{\,2}}(y - \overline{Y}) + \overline{X}$$

Ces équations permettent de définir deux droites différentes de régression à l'intérieur du nuage de point. Néanmoins cette inversion, qui permet d'obtenir l'équation x= a'y + b' (régression de x en y) n'est pas souvent intéressante, car en général, Y est une variable à exprimer et X est une variable potentiellement explicative.

Propriété de ces deux droites de régression :

1) les deux droites de régression se coupent en un point qui a pour coordonnées les moyennes de x et de y, point $(\bar{x}, \bar{y})$, (en remplaçant dans l'équation x par sa moyenne, il est ainsi possible de retrouver y (qui correspond à la moyenne de y)).

2) les coefficients a et a' (qui sont les pentes) sont toujours de même signe (soit – (corrélation négative) soit + (corrélation positive)), ainsi les deux droites sont orientées dans le même sens que le nuage de point.

3) l'angle maximum des deux droites de régression est de 90° (droites perpendiculaires). Dans ce cas, les points sont dispersés dans tout le plan. La corrélation est nulle. Les droites sont respectivement parallèles à l'axe des x et à l'axe des y.

Remarque : Les « fausses corrélations »

Qu'est-ce qu'une corrélation ? C'est une relation positive ou négative entre deux phénomènes, mais elle n'est pas absolue. Ainsi, il y a une corrélation positive entre la taille et le poids des hommes : ceux qui mesurent un mètre quatre-vingt pèsent en général plus lourd que ceux dont la taille ne dépasse pas un mètre soixante. Mais il y a des petits gros et des grands maigres.

Souvent, une corrélation est le signe d'une relation de cause à effet. Le plus souvent, on sait ce qui est la cause et ce qui est l'effet : c'est la consommation de tabac qui provoque le cancer du poumon et non la prédisposition à ce cancer qui donne envie de fumer. Mais dans certains cas, les choses sont beaucoup moins évidentes. Et il peut arriver aussi que chacun des deux phénomènes soit à la fois cause et effet.

En outre, il y a beaucoup de corrélations statistiques qui ne résultent aucunement d'une relation de cause à effet et qui sont de ce fait trompeuses. C'est notamment le cas pour les séries statistiques qui évoluent parallèlement dans le temps, avec le progrès économique et scientifique. Certes, si l'espérance de vie augmente, en même temps que diminue la fréquentation des cinémas (corrélation négative), personne n'ira soutenir que l'on vit plus vieux parce que l'on va moins souvent au cinéma. Mais dans bien des cas, surtout si l'on veut prouver quelque chose, on n'hésitera pas à voir une relation de cause à effet là où il n'y a rien d'autre que l'évolution parallèle de deux séries statistiques.

APPLICATION IX: AJUSTEMENT LINÉAIRE

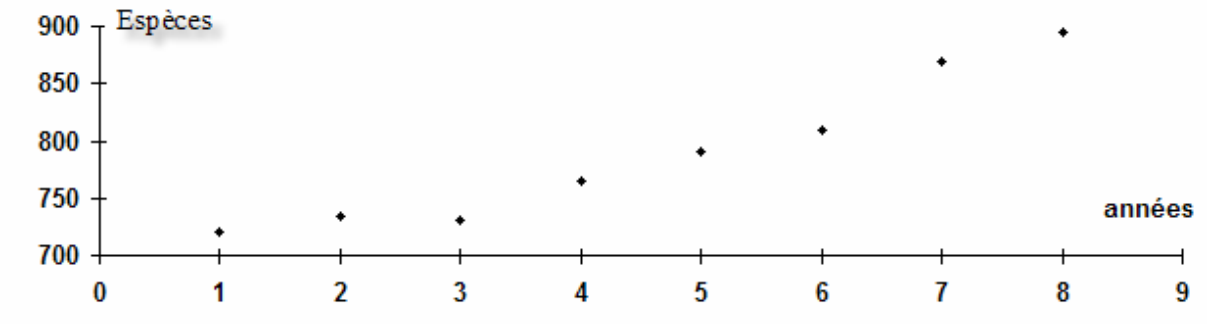
I. Savoir représenter graphiquement une série chronologique :

- **Exemple :** On référence sur huit années, le nombre d'espèces affectées par une substance toxique

Années	01	02	03	04	05	06	07	08
Nb d'espèces	720	735	730	765	790	810	870	895

Représenter graphiquement cette série.

- **Méthode :** on porte en abscisse les numéros des années, $x_i (1 \le x_i \le 8)$ et respectivement en ordonnées les effectifs des espèces affectées par cette toxine.



II. Ajuster la série chronologique par la méthode des points moyens:

- **Exemple :** ajuster la série précédente à l'aide d'une droite en utilisant la méthode des points moyens.
- **Méthode :** on détermine l'équation de la droite passant par deux points moyens *A* et *B*.

A a pour abscisse la moyenne des abscisses correspondant des années 1 à 4 et pour ordonnée la moyenne des espèces. De même *B* pour les années 5 à 8.

- **Solution :** soient $A(x_A; y_A)$ et $B(x_B; y_B)$ les deux points moyens.

$$x_A = \frac{1+2+\cdots+4}{4} = 2{,}5 \qquad y_A = \frac{720+735+730+765}{4} = 737{,}5$$

$$x_B = \frac{5+6+\cdots+8}{4} = 6{,}5 \qquad y_B = \frac{790+810+870+895}{4} = 841{,}25$$

Le coefficient directeur de la droite d'équation générale $y = ax + b$ est :

$$a = \frac{y_B - y_A}{x_B - x_A} = \frac{841{,}25 - 737{,}5}{6{,}5 - 2{,}5} \approx 25{,}9$$

Sur *A* on a : $y_A = ax_A + b \Rightarrow 737{,}5 = 25{,}9 \times 2{,}5 + b$

D'où $b = 737{,}5 - 25{,}9 \times 2{,}5 = 672{,}75$

La droite a donc pour équation : $y = 25{,}9x + 672{,}75$

Cela permet d'effectuer des prévisions. Par exemple, pour l'année n° 9, le nombre d'espèces affectés prévisionnel sera de : $y_9 = 25{,}9 \times 9 + 672{,}75 = 905{,}85$ (soit environ 906 espèces)

III. Ajuster la série par la méthode des moindres carrés :

- **Exemple :** ajuster la série précédente à l'aide d'une droite en utilisant la méthode moindres carrés.
- **Méthode :** on commence par calculer la moyenne $\bar{x}$ des années (x_i) et la moyenne $\bar{y}$ des espèces (y_i).

On obtient le point moyen *M* de la série, point par lequel passe la droite d'ajustement.

On calcule ensuite le coefficient directeur de la droite : $a = \dfrac{\sum X_i Y_i}{\sum X_i^2}$ avec $X_i = x_i - \bar{x}$ et $Y_i = y_i - \bar{y}$ (écarts par rapport aux moyennes respectives).

- **Solution :** on construit le tableau de calculs suivants :

Années xi	Espèces yi	Xi	Yi	XiYi
1	720	- 3,5	-69,38	242,81
2	735	- 2,5	-54,38	135,94
3	730	- 1,5	-59,38	89,06
4	765	- 0,5	-24,38	12,19
5	790	0,5	0,63	0,31
6	810	1,5	20,63	30,94
7	870	2,5	80,63	201,56
8	895	3,5	105,63	369,69
36	6 315			1 082,50

On a : $\bar{x} = \dfrac{36}{8} = 4{,}5$ et $\bar{y} = \dfrac{6315}{8} = 789{,}375$ coordonnées du point moyen M.

Coefficient directeur de la droite : $a = \dfrac{1082{,}50}{42} \approx 25{,}77$

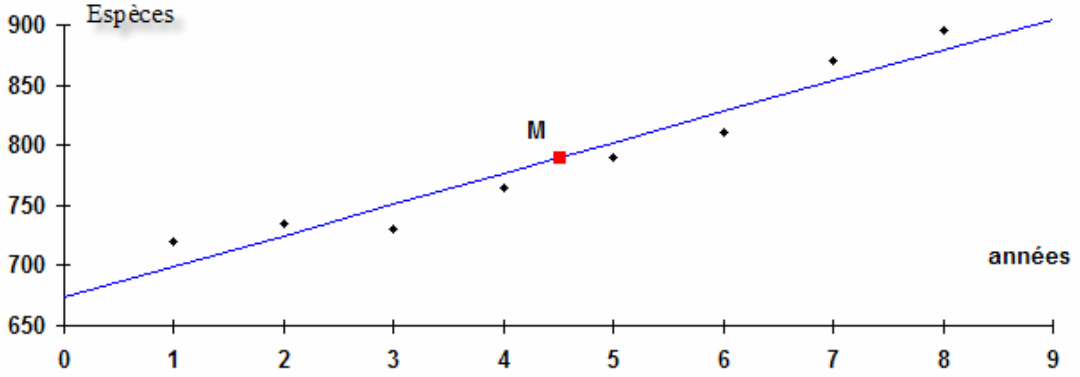
Sur *M* on a : $\bar{y} = a\bar{x} + b \Rightarrow 789{,}375 = 25{,}77 \times 4{,}5 + b$

D'où $b = 789{,}375 - 25{,}77 \times 4{,}5 = 673{,}41$

La droite a donc pour équation : $y = 25{,}77x + 673{,}41$

Cela permet d'effectuer des prévisions. Par exemple, pour l'année n° 9, le CA prévisionnel sera de : $y_9 = 25{,}77 \times 9 + 673{,}41 = 906{,}34$ espèces (environ 907 espèces) légèrement différente de la méthode des points moyens.

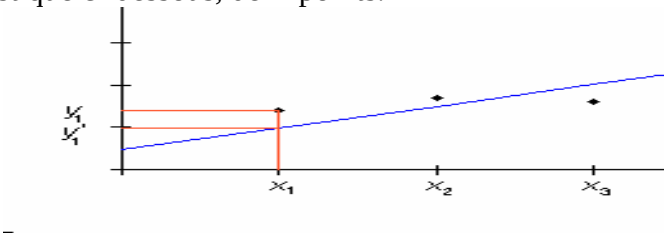
- **Graphique :**



IV. Méthode des moindres carrés : démonstration

On dispose de la série statistique ci-dessous, de *N* points.

x	y
x_1	y_1
x_2	y_2
...	...
x_N	y_N



On suppose que la forme du " nuage de points " permet d'envisager légitimement un ajustement linéaire à l'aide d'une droite.
La droite a une équation de la forme $y' = ax + b$. Pour chaque x_i, on va chercher à minimiser les carrés des écarts entre les ordonnées y_i du point correspondant de la série et les ordonnées des points de la droite $y'_i = ax_i + b$.

Le problème consiste donc à déterminer les coefficients *a* et *b* tels que si on calcule $y'_i = ax_i + b$, on doit rendre minimum l'expression : $\sum_1^N (y_i - y'_i)^2 = \sum_1^N (y_i - ax_i - b)^2$

Supposons *a* connu, donc fixé. Seul b " bouge ". La fonction du 2ème degré $f(b) = \sum (y_i - ax_i - b)^2$ passe par un minimum quand la dérivée par rapport à *b* s'annule

$$\left(\sum (y_i - ax_i - b)^2\right)' = 0 \Leftrightarrow -2\sum (y_i - ax_i - b) = 0 \Leftrightarrow \sum y_i - a\sum x_i - Nb = 0$$

D'où : $b = \frac{\sum y_i}{N} - a\frac{\sum x_i}{N} \Rightarrow b = \bar{y} - a\bar{x}$ avec $\bar{y} = \frac{\sum y_i}{N}$ (moyenne des y_i) et $\bar{x} = \frac{\sum x_i}{N}$ (moyenne des x_i). Autrement dit, la fonction $y' = ax + b$ dont nous cherchons les coefficients est telle que $\bar{y} = a\bar{x} + b$, la droite d'ajustement passe donc par le point moyen $M(\bar{x}, \bar{y})$. On a donc déterminé *b* en fonction de *a*.

L'équation cherchée devient $y' = ax + \bar{y} - a\bar{x} \Leftrightarrow y' - \bar{y} = a(x - \bar{x})$

Posons $Y = y' - \bar{y}$ et $X = x - \bar{x}$ ce qui revient à prendre $M(\bar{x}, \bar{y})$ comme nouvelle origine.
Notre droite d'ajustement passant par cette origine aura une équation de la forme $Y = aX$.
Dans ce nouveau repère, les points de notre série statistique auront pour coordonnées :
$X_i = x_i - \bar{x}$ et $Y_i = y_i - \bar{y}$. Pour obtenir *a*, on doit minimiser l'expression

$$\sum_1^N (y_i - ax_i - b)^2 = \sum_1^N (y_i - ax_i - (\bar{y} - a\bar{x}))^2 = \sum_1^N (y_i - \bar{y} - a(x_i - \bar{x}))^2 = \sum (Y_i - aX_i)^2 \quad .$$

Seul *a* " bouge " et est variable. La fonction du 2ème degré $g(a) = \sum (Y_i - aX_i)^2$ passe par un minimum lorsque la dérivée par rapport à *a* s'annule.

$$\left(\sum (Y_i - aX_i)^2\right)' = 0 \Leftrightarrow -2\sum (Y_i - aX_i) X_i = 0 \Leftrightarrow \sum X_i Y_i - a\sum X_i^2 = 0$$

D'où l'on tire :

$$a = \frac{\sum X_i Y_i}{\sum X_i^2}$$

CHAPITRE III.

INFORMATIQUE ET STATISTIQUE :
Pré-requis, mise à niveau et apprentissages

1. INFORMATIQUE : PRE-REQUIS ET MISE A NIVEAU

1.1. Matériels et interfaces utiles (disponible au laboratoire d'informatique)
- Calculatrice scientifique (possible d'utiliser celle du PC ou celles fournis)
- PC interface avec office 2003 et plus (Word, Excel,..),
- Autres logiciels et plug-in: adobe acrobate reader (lecture et impression des fichiers PDF), flash player (plug-in pour navigateur et lecteur d'animations SWF (Animations **Flash**)), compression/décompression (winzip et winrar), Exlstat (outil d'analyse de données et de statistiques pour Microsoft Excel), lecteur de format et fichier « chm » ;

1.2 Pré requis
- Interface Windows (de préférence XP ou Vista)
- Interface et fichier Word
- Interface et fichier Excel
- Interface et fichier PDF
- Gestion d'images et de format d'images (jpg, gif, png,...)
- Notion d'Internet (navigation, téléchargement, mailing, forum,…)
- Gestion de fichiers compressés (Winzip, winrar,..)

1.3 Mise à niveau théorique et pratique
- Mise à niveau des pré-requis

2. APPRENTISSAGES INFORMATIQUE ORIENTE STATISTIQUE
2.1. Gestion de données numérique et de tableau sur Word et Excel
2.2. Gestion de calculs et de formules statistique dans Excel
2.3. Gestion et élaboration de calcul statistique sur Excel
2.4. Gestion et élaboration de représentations graphiques sur Excel
2.5. Utilisation et insertion de Macro dans Excel
2.6. Représentation de séries et calculs statistique (tableau et graphique)
2.7. Ajustement linéaire de séries chronologiques avec et sans variations saisonnières.
2.8. Présentation et principe de logiciel d'analyse de donnée statistique
2.9. Utilisation Excel et présentation d' XLSTAT comme outil pour statistique descriptive

APPLICATIONS ET TRAVAUX DIRIGES

EXERCICES APPLIQUES : STATISTIQUES ET INFORMATIQUE

OBJECTIF

Ces travaux dirigés sont l'occasion d'illustrer à travers des simulations et des exemples concrets les différentes notions de base de statistique vues en cours. Nous insisterons sur les résultats fournis par les logiciels statistiques car ceux-ci sont de plus en plus élaborés, et les résultats qu'ils fournissent sont plus ou moins facilement interprétables. Le logiciel que nous utiliserons tout le long de ces travaux dirigés est Excel et Exstat.

- TD 1 - INITIATION A LA MANIPULATION D'UN TABLEUR (Notion d'informatique) (Tableur Excel)

Activité I : Mise à niveau des pré-requis

Exercice 1. Evaluation des pré-requis et mise en pratique

Activité II. Gestion de données numériques et de tableau sur Word et Excel

Exercice 1 : Construction de table,
Exercice 2 : Conversion, transformation et ajustement de données alphanumériques
Exercice 3 : Trie, sélection, substitution de données alphanumériques

Activité III : Utilisation de feuille de calcul Excel

Exercice 1 : Construction de table de multiplication dynamique
(Utiliser le tutoriel (table_multiplication.swf) en flash)
Exercice 2 : Fonctions de bases d'Excel
Exploitation de la fonction formule

Activité IV : Excel : Construction graphique et application statistique simple

Exercice 1 : Construction d'un histogramme
Voir tutoriel: Exemple histogramme 3D sur Excel.avi
Exercice 2: Construction de courbes

TD2 : Activités : Initiation aux utilitaires mathématique et statistique de l'interface Microsoft

A. Quelques utilitaires de l'interface Microsoft

1- La calculatrice scientifique de l'interface Microsoft
2- L'insertion de symbole de Microsoft Word
3- L'éditeur d'équation de Microsoft Word

Activité A1: Calculez avec la calculatrice scientifique la somme, la moyenne et l'écart-type de la série statistique suivante : 20, 21, 21, 19, 23, 22

B. Quelques fonctions usuelles d'Excel

Rechercher dans l'aide de Microsoft Excel :
1- les fonctions mathématiques et trigonométriques
3- quelques astuces sur Excel (raccourcis claviers, notion de macro...)
2- les fonctions statistiques

C. Application et utilisation de fonctions statistiques sur Excel

1- Passage de données en vrac aux données rangées et aux données condensées.
2- Elaboration de données en classes
3- Tableau des fréquences
4- Tableau de distribution
5- Elaboration et gestion des graphiques
6- Gestion des calculs
Sommes, Moyenne, Mode, Médiane, Quartile, Décile, Centiles, Maximal, Minimal
Fréquences, Variance, Ecart-type, Covariance, Coefficient de variation

D. Construction de représentations graphiques élaborées

1- Construction de séries chronologiques
2- Construction de Pyramide
3- Construction et exploitation de barres à moustaches

TD3: Activités questions de synthèses :

1- Enumérer à travers le fascicule l'ensemble des questions qui peuvent être posées.

- Enoncer les formules respectives des paramètres suivants :
Moyenne, Moyenne pondéré, la variance et l'Ecart-type avec N>30 et N<30,
- Calculer : les fréquences relatives et absolues
- Calculer : la moyenne, le mode, la médiane, la variance, l'écart-type,
- Calculer : les effectifs cumulés croissant et décroissant
- Calculer la moyenne, la variance et l'écart-type en utilisant la méthode des ζ
- Trouver la valeur Xi du rang décile Cα
- Trouver le rang décile Cα de la valeur Xi
- Elaborer un tableau détaillé de distribution, en indiquant respectivement les valeurs xi, mi, ni, fi et Fi.
- Tracer un histogramme des effectifs
- Tracer un histogramme des fréquences
- Tracer la courbe des fréquences cumulées croissante et décroissante, en déduire la médiane
- Tracer des barres à moustaches et comprendre leurs indicateurs
- Ajuster une série chronologique par les méthodes des points moyens et des moindres carrés

PLANCHE D'ACTIVITES

Pour toutes les activités, vous devez formuler un ensemble de questions, puis analyser la situation en choisissant la procédure et les tests statistiques appropriés. La correction doit être fournie sur document Word et sur document Excel. Sur le document indiquez : Nom, Prénom, N° d'inscription

TABLEAU A

Adultes de 18 à 20 ans	Triglycéride
1	152
2	59
3	117
4	54
5	93
6	176
7	79
8	89
9	307
10	88
11	299
12	52
13	158
14	98
15	101
16	71
17	81
18	86
19	71
20	71
21	107
22	80
23	47
24	95
25	140
26	77
27	57
28	95
29	480
30	94

TABLEAU B

Taux de cholestérol (30 étudiants)

197	194	137	215	212
181	155	285	194	175
190	234	218	207	158
131	201	167	198	115
172	258	170	189	228
233	212	157	216	164

TABLEAU C

Périmètre du tronc	Poids de l'arbre
358	760
375	821
393	928
394	1009
360	766
351	726
398	1209
362	750
409	1036
406	1094
487	1635
498	1517
438	1197
465	1244
469	1495
440	1026
376	912
444	1398
438	1197
467	1613
448	1475
478	1571
457	1506
456	1458
389	944
405	1241
405	1023
392	1067
327	693
395	1085
427	1242
385	1017
404	1084
416	1151
479	1381

ACTIVITE 1 : Le tableau A présente les résultats d'un test de dépistage de triglycérides chez des adultes de 18 à 20 ans. Avant de faire une analyse statistique, regrouper ces résultats sous forme de données condensées.

ACTIVITE 2 : Lors d'un examen médical, on a voulu mettre en place un dépistage de lipides sanguins chez une promotion d'étudiant en médecine. Le but étant de s'assurer que le taux de cholestérol moyen des étudiants est inférieur à 190 (taux au dessus duquel le cholestérol peut être nocif). Les données sont résumées dans le tableau B.

ACTIVITE 3 : Les données du tableau C proviennent des archives d'un laboratoire de recherche. Elles renseignent sur un suivi (de plus de 30 années) de l'évolution des arbres d'une réserve naturelle. Ces données nous permettent d'obtenir des estimations de poids sans avoir à couper les arbres pour les peser, méthode destructive et problématique. Faites une analyse des relations entre ces 2 variables. Peut on estimer de façon précise le poids d'un arbre dont le périmètre est égal à 525 ?

QUELQUES STATISTICIENS

Pour une histoire de la statistique. Tome 1. Insee, Imprimerie Nationale, 593 pp.

- ACHENWALL Gottfried
Juriste allemand (1719 - 1772). Professeur de droit international et de science politique à Goettingue, il diffusa le mot "statistique". Il emprunta ce mot à Marton Schmeizel, qui fut son professeur et qui était lui-même élève de Conring.

- ARBUTHNOT J.
Médecin et écrivain écossais (Arbuthnot, Kincardineshire, 1667 - Londres, 1735). Il remarqua que, parmi les enfants baptisés à Londres chaque année de 1629 à 1710, le nombres des garçons dépassait toujours celui des filles. Considérant que celà prouvait que les probabilités, pour chaque naissance, d'obtenir un garçon ou une fille n'étaient pas égales, il attribua cette inégalité à la "Divine Providence". Certains font remonter les méthodes non paramétriques à cette observation. 1710. An argument for Divine Providence, taken from the constant regularity observ'd in the births of sexes. *Phil. Trans. R. Soc*, 27: 186-190.

- ARISTOTE (en grec Aristotelês, dit le Stagirite)
Philosophe grec (Stagire, Macédoine, aujourd'hui Stavro, - 384 - Chalcis, Eubée, -322). L'oeuvre d'Aristote comporte également des traités de politique (*Politique*, où apparaît l'origine des statistiques descriptives qui se répandront en Europe au XVIe siècle; *Constitution d'Athènes*)

- ***ARTHASASTRA***
Traité de science politique et économique rédigé par Kautilya, ministre du roi Candragupta du premier Empire indien des Maurya (IVe siècle avant notre ère). Il est remarquable, entre autre, par la description des techniques perfectionnées de recensement de la population et de statistiques.

- BODIN Jean
Economiste et philosophe français (Angers, 1530 - Laon, 1596). Dans son traité *Methodus ad facilem historiarum cognitionem*, il a montré l'importance de la connaissance de l'histoire pour la compréhension du droit et de la politique. Economiste, il a analysé le phénomène de la montée des prix au XVIe siècle en relation avec l'apport des métaux précieux d'Amérique (*Réponse aux paradoxes de Malestroit*).Dans son traité de science politique (*La République*, 1576), théorie de la monarchie absolue, il démontre l'intérêt de l'idée de dénombrement, base de la statistique descriptive. Cette idée connaîtra un grand succès et sera reprise, plagiée sans que les emprunteurs citent leur source; ainsi le *Miroir des François* de N. de Montand (1581) et le *Traité de l'économie politique* d'Antoine de Montchrétien (1615).

- BURT sir Cyril
Psychologue britannique (Londres, 1883 - 1971). Il fut le psychologue officiel du London County Council, responsable de l'application et de l'interprétation des tests mentaux dans les écoles de Londres. Il succéda à Charles spearman à la chaire de l'University College de Londres (1932 - 1950). Spécialiste de la statistique pyschologique, il réifia l'analyse factorielle en assimilant un axe factoriel mathématique au concept d'"intelligence générale", à la suite de C. Spearman. Il considérait que l'intelligence, concept nébuleux, pouvait s'identifier à une "chose" possédant une localisation précise dans le cerveau et un degré d'héritabilité, chose que l'on pouvait mesurer et réduire à un chiffre permettant de classer les individus en fonction de la quantité qu'ils en possèdent. Sa théorie héréditariste de l'intelligence était basée sur l'étude de couples de jumeaux vrais (une cinquantaine) élevés séparement dont les QI étaient en forte corrélation. Il fut célèbre et couvert d'honneurs. Il fut démontré, quelques années après sa mort, qu'il était l'auteur d'une gigantesque supercherie scientifique. Il avait inventé ses couples de jumeaux, sa collaboratrice et ses résultats. *Mental and scolastic tests*. London Country Council, 432 p., 1921; *The backward child*, Appleton,

New York, 694 p., 1937; *The factors of mind*, Univ. London Press, 509 p., 1940; *Intelligence and fertility*, Eugenics Society, 43 p., 1946; *The Causes and treatment of Backwardness*, 1952.

- CONRING Hermann
Juriste allemand (1606 - 1632). Professeur de droit public à Helmstedt, il introduisit pour la première fois l'enseignement de la statistique à l'Université. Ses notes de cours (*Examen rerum publicarum potiorum totius urbis*) furent publié en 1667. Il y décrit de nombreux pays européens et non européens, sans apporter de données chiffrées. La statistique est la science de la constitution de l'état, mais elle est purement descriptive.

- FISHER Ronald
Statisticien anglais (1890 - 1962). À partir de ses expérimentations agronomiques, il tenta de montrer que, même si les postulats de normalité relevaient souvent de l'abus de confiance, cela ne détériorait pas trop la validité des conclusions. Dans ce cadre, il fut l'un des premiers à développer les tests de permutations des rangs ou de randomisation avec Pitman et Welch. *The design of experiments*. Oliver & Boyd, Edimbourg, 1935.

- GALTON sir Francis
Physiologiste, anthropologue et psychologue anglais (Birmingham, 1822 - Haslemere, 1911). Cousin de C. Darwin, il fut un pionnier des statistiques modernes. Il pensait que la mensuration était le critère primordial de toute étude scientifique. Il entreprit même une enquête statistique sur l'efficacité de la prière. Sa croyance, que même les comportements les plus enracinés dans la société avaient une composante innée, l'amena à étudier l'hérédité et les différences individuelles (héréditarisme; *Hereditary Genuis*, 1869; *English men of science, their nature and nurture*, 1874; *Inquires into Human Faculty and its Development*, 1883; *Natural Inheritance*, 1889). L'anthropométrie et le mesurage des crânes et des corps furent parmi les critères les plus utilisés. Il étudia le niveau d'intelligence d'un individu à l'aide de l'étalonnage des tests. Il fut un des fondateurs de "l'eugénisme" (*Essays on Eugenics*, 1909). Mais, il fut un des premiers à réaliser que les valeurs moyennes attachées à des populations biologiques pleines de variabilité ne sont que des artifices de calcul.

- KAUTILYA
Ministre du roi Candragupta, fondateur de la dynastie et du premier Empire indien des Maurya (313 - 226). Il rédigea un traité de science politique et d'économie. Il justifie le recours aux recensements, à la statistique et au cadastre pour remplir son rôle de planificateur. Il montre l'intérêt porté par les empires asiatiques (populationniste) au dénombrement de leur population.

- MORTON Samuel George
Médecin et anthropologue américain (Philadelphie, 1799 - 1851). Il collectionna jusqu'à sa mort plus d'un millier de crânes humains et fut reconnu comme le premier objectiviste de la science américaine. Cette collection était réalisée en vue d'établir une classification objective des races humaines en se fondant sur les caractères physiques du cerveau et de sa taille. Il fournit et analysa les données permettant de soutenir le polygénisme et devint un des dirigeants du mouvement polygéniste américain. Il publia toutes ses données brutes, mais ses résumés sont un ramassis d'erreurs de calcul, d'astuces et de tripotages de chiffres, inconscients, tendant à confirmer ses convictions préalables (voir S.J. Gould). *Crania americana* (1839) est un traité sur l'infériorité de l'intelligence chez les indiens, contenant des erreurs statistiques et un échantillon biaisé lui permettant de "calculer" une moyenne faible pour les cerveau des indiens, nettement inférieure à celle des blancs. Le *Crania Aegyptiaca* (1844) établit à partir d'un échantillon de 100 crânes provenant des catacombes de l'ancienne Egypte lui permit de parfaire sa thèse en démontrant que l'écart séparant les blancs des noirs était encore plus grand. On y retrouve les nombreuses petites erreurs numériques du premier livre et l'erreur, plus fondamentale, de la non liaison entre capacité crânienne et taille, sans compter la variabilité introduite par sa procédure de mesure de la capacité crânienne.

- PEARSON Karl
Statisticien anglais (1857 - 1936). "Il s'est donné pour mission de faire fructifier au niveau théorique les problèmes posés par l'application de la statistique à la biologie : il se consacre donc à l'étude des probabilités, mettant au point la fameuse formule du chi carré". F. Bédarida, 1977. Statistique et société en Angleterre au XIX e siècle. In *Pour une histoire de la statistique*, INSEE. Il soutint Galton et sa théorie de "l'eugénisme".

- PITMAN EJG.
Il fut l'un des premiers à développer les tests de permutations des rangs ou de randomisation avec Fisher et Welch. Significance tests that may be applied to samples from any population. *J. R. Stat. Soc., Suppl.*, 4: 119-130, 1937. Significance tests that may be applied to samples from any population. III. The analysis of variance test. *Biometrika*, 29: 322-335, 1938.

- PLAYFAIR William
XIXe siècle. Voyageur, dessinateur, statisticien, économiste, inventeur. Il inventa la méthode d'expression des faits statistiques par des procédés géométriques. La première illustration de la méthode graphique a été donné dans *Commercial and political atlas* 1786. Il présente deux types de graphiques : des courbes et des histogrammes. Le nom de ce dernier a été inventé par Pearson en 1895. Il présente les graphes circulaires et les diagrammes à sections dans *Statistical breviary* 1801.

- QUETELET Adolphe
Statisticien belge (1796 -1874). Il fut un disciple de Laplace. Il recherchait des lois déterministes et espérait calculer les caractéristiques de "l'homme moyen", c'est-à-dire de découvrir l'essence (le type) de l'homme. Les variations n'étaient que des "erreurs" autour de la moyenne. Il joua un rôle éminent dans la création de la statistique mathématique.

- SPEARMAN Charles
Psychologue et statisticien anglais (Londres, 1863 - Londres, 1945). Fondateur de la psychologie différentielle, il mis au point la méthode mathématique de l'analyse factorielle (1904). Il admit que la réussite à une tâche (test) est déterminée par une aptitude générale, le facteur g (intelligence globale) intervenant dans toutes les épreuves psychologiques et une aptitude spécifique à la tâche particulière. Il justifia, d'un point de vue théorique, l'usage d'une échelle linéaire de Q.I., que Binet avait proposé comme un simple guide empirique, sur l'analyse factorielle elle-même. Il s'enferma dans de profondes erreurs conceptuelles dont la principale fut le réification de l'intelligence. Il identifia un concept nébuleux, socialement défini, comme l'intelligence à une "chose" possédant une localisation précise dans le cerveau et un degré d'héritabilité. Il était alors possible de mesurer cette chose et de la réduire à un chiffre unique permettant de classer les individus en fonction de la quantité qu'ils en possèdent (Q.I.).
Il proposa le coefficient de corrélation de rangs qui porte son nom, premier test de statistique nonparamétrique.
The proof and measurement of association between two things. *Am. J. Psychol*., 15: 72-101, 1904. *The nature of "intelligence" and the principles of cognition*. Londres, McMillan, 358p., 1923; *Les aptitudes de l'homme. Leur nature et leur mesure*. McMillan, Londres, 1927; *Psychology down the ages*. McMillan, Londres, 2 vol., 454 et 355 p., 1937; Spearman C. & J. L. Wynn : *Human ability*, McMillan, 198 p., Londres, 1950.

- STUDENT (GOSSET) William Sealy
Statisticien anglais (1876 - 1937). W.S. Gosset a publié sous le nom de Student. Il travaillait pour l'industrie de la bière (maison Guinness). Il fit progresser la statistique dans le domaine des probabilités.

LEXIQUE FRANÇAIS / ANGLAIS

Ecart-type ou écart quadratique moyen ou déviation standard : *Standard deviation*
échantillon : *Sample*
échantillonnage : *Sampling*
Erreur de seconde espèce : *Second kind error, bêta-error*.
Erreur de première espèce : *First kind error, alpha-error*
Hypothèse alternative (H_1) : *Non-null hypothesis ou Alternative hypothesis*
Hypothèse nulle (H_0) : *Null hypothesis*
Homoscédasticité : *Homoscedasticity*
Niveau de signification (alpha) : *Significance level*
Population : *Population*
Pouvoir d'un test : *Test power*
Région d'acceptation ou de non-rejet : *Acceptance region*
Région de rejet ou domaine de rejet ou région critique: *Rejection region*
Seuil de signification ou valeur critique : *Significant point ou Critical value*
Tests d'ajustement : *Test of goodness of fit*
Test binomial : *Binomial test*
Test d'hypothèses ou tests de signification : *Test of hypothesis ou Significance tests*
Tests statistiques bilatéraux :*Double-tailed test* ou *two-sided test*
Tests unilatéraux : *Single-tailed test* ou *one-sided test*

***Acceptance region* :** Région d'acceptation ou de non-rejet
***Binomial test* :** Test binomial
Double-tailed test or two-sided test : Tests statistiques bilatéraux
Double-tailed test or two-sided test *:* Tests statistiques bilatéraux
First kind error, alpha-error : Erreur de première espèce
Homoscedasticity : Homoscédasticité
***Non-null hypothesis or Alternative hypothesi* s :** Hypothèse alternative (H1*)*
***Null hypothesis* : Hypothèse nulle (H0)**
***Rejection region or critical region* :** Région de rejet
Sample : échantillon
Sampling : échantillonnage
Second kind error, bêta-error . : Erreur de seconde espèce
***Significance level* :** Niveau de signification (alpha)
Significant point or Critical value : Seuil de signification ou valeur critique
Standard deviation : Ecart-type ou écart quadratique moyen ou déviation standard
Test of goodness of fit : Tests d'ajustement
Test of hypothesis ou Significance tests : Test d'hypothèses ou tests de signification
Test power : Pouvoir d'un test

Bibliographie

BERTIN J. 1977. *La graphique et le traitement graphique de l'information.* Nouvelle bibliothèque scientifique, Flammarion.

CAPERAA Philippe & VAN CUTSEM Bernard, 1988. *Méthodes et modèles en statistique non paramétrique. Exposé fondamental*. Presses Université Laval, Dunod, 357 pp.

DAGNELIE Pierre, 1969 - 1970. *Théorie et méthodes statistiques. Applications agronomiques* (3 vol.). Duculot, Gembloux, Presses Agron., 378 + 451 pp.

FISHER R.A., 1946.*Statistical methods for research workers*, Olivier & Boyd, London. traduction française aux Presses Universitaires.

HAYS W. L., 1963. *Statistics for psychologists*. Holt, Rinehart & Winston.

LE GUELTE L., LE BERRE M., DAHAN G., RAMOUSSE R. & COULON J. 1983. Traitement statitistique informatisé des données en éthologie. *Études et analyses comportementales,* 1(4) :202-268.

Pour une histoire de la statistique. Tome 1. Insee, Imprimerie Nationale, 593 pp.

SCHWARTZ D. 1963. *Méthodes statistiques à l'usage des médecins et des biologistes.* Paris, Flammarion Médecine Sciences. SIEGEL Sidney, 1956. *Non parametric statistics for the behavioral sciences* , McGraw Hill, 312 pp.

SNEDECOR G.W. *Calculation and interpretation of analysis of variance and covariance*, Collegiate Press, Ames, Iowa.

SPRENT P. 1992. *Pratique des statistiques nonparmétriques.* INRA Editions. VESSEREAU A. 1948. *Méthodes statistiques en biologie et en agronomie*. Baillère et fils, Paris, p.381.

VIGNERON E. 1997. *Géographie et statistique.* Que sais-je?, PUF. WINER B.J. 1970. *Statistical principles in experimental design.* McGraw-Hill, Mladinska Knijiga, p. 672.