Important : à l'examen, seules les calculatrices sans mémoire possédant les opérations (+ - / racine, puissance) sont autorisées.

COURS DE STATISTIQUES

I. INTRODUCTION AU COURS

La statistique a été mise au point pour traiter en général de grands ensembles de données numériques, comme celles qui peuvent être acquises lors d'enquêtes, de sondages ou par des appareils de mesure. La géographie, comme d'autres disciplines, est confrontée à d'énormes masses de données : en géographie humaine (chiffres économiques, populations, indices de croissance) en géographie physique (échantillons de sol, d'eau…), où parfois à l'aide d'appareils installés en permanence et qui donnent de grandes séries de mesures (Températures, débits des cours d'eau, niveau d'une nappe phréatique...). Toutes ces données peuvent évoluer dans le temps, mais aussi dans l'espace, ce qui complique leur analyse.
La statistique va nous offrir une certaine image de la réalité grâce à de multiples techniques qui ont leurs avantages ou leurs inconvénients et, il faut le savoir aussi, qui sont parfois très critiquées. Ces méthodes ont un pouvoir descriptif et explicatif bien plus riche que si on se contente d'interpréter des tableaux élémentaires de données, elles permettent de mettre en évidence ce qui est essentiel dans les données .
En effet, la statistique permet de trouver des liens entre les variables et de dégager des structures dans les données, de trouver des renseignements pertinents noyés dans une masse d'information, elle peut également classer l'information : elle synthétise l'information.
Pour réussir à bien déchiffrer ces renseignements, pour éviter de se tromper lors d'une analyse de données, pour comprendre la signification de certains chiffres, pour employer la bonne méthode d'analyse, pour comprendre la puissance des méthodes statistiques, il faut maîtriser les notions fondamentales de la statistique.
Définition : la statistique est une famille de techniques visant au rassemblement, à la présentation et à l'analyse des données quantitatives, ainsi qu'à l'utilisation de ces données pour prendre des décisions.
Au départ, la statistique se contentait de comprendre les problèmes en étudiant le passé et en faisant plutôt des analyses rétrospectives et historiques. Aujourd'hui on fait de la statistique pour anticiper l'avenir : on prévoit des expériences, on choisit des échantillons, on analyses des données en fonction de décisions qui doivent être prises et elles mènent à des actions concrètes dans le domaine politique et surtout économique (études de clientèle, marketing, études des goûts et habitudes…).

La statistique comporte deux niveaux :

1) La statistique descriptive

Elle consiste à rassembler les données de base à les mettre en forme dans des tableaux, les critiquer (détecter des valeurs aberrantes, anormales, hors norme et prendre la décision de les garder où non), les présenter et en faire une première analyse sommaire CAD : les grandes masses de données doivent être condensées, résumées, pour qu'on puisse facilement en prendre connaissance.

2) La statistique exploratoire

Tente d'élaborer des théories et d'énoncer des lois, voir d'extrapoler les résultats pour éventuellement anticiper l'avenir, grâce à des modèles à bases probabilistes.
Nous allons aborder ensemble :

1) Révisions et notions générales de statistique descriptive
2) Les indices statistiques (servent à mesurer l'évolution au cours du temps de phénomènes économiques) ;

Dans la deuxième partie, vous aborderez quelques notions complémentaires comme la corrélation ou l'étude des séries temporelles.

II. REVISIONS ET BASES

A. Conventions d'écriture

Ces conventions simplifient l'écriture des formules et la plupart d'entre vous les connaissent certainement déjà.

1) Constante/variable

Une constante est souvent désignée par une lettre minuscule (a, b ou c).©

Une variable est notée X, Y, Z©. Une variable sera capable de prendre toute une série de valeurs.

X	1	2	5	1	8	9

2) Rang/indice

Le rang s'est la place qu'occupe une valeur dans la série : la valeur qui occupe le rang 3 c'est 5.

Il y a autant de rangs possibles dans la série que de valeurs. En général N ou n© est la lettre qui désigne le nombre total de rangs possibles. Dans la série ci-dessus, N vaut 6.
En statistique on a souvent besoin de désigner le rang des valeurs, s'est pourquoi on utilise l'écriture indicielle. Les indices sont placés généralement en bas et à droite des lettres. Ils désignent le rang de l'une des valeurs de la série. Ces indices sont soit des lettres (ils désignent n'importe quel rang), soit des chiffres, ils désignent un rang précis.
Exemples :

L'écriture d'une valeur x de rang quelconque i (se lit « x indice i ») : x_i©
i est un indice qui varie de 1 à N.

Ecriture d'une valeur x de rang 3 : x_3© et $x_3 = 5$

3) Somme

Pour une variable X composée de N éléments, on cherche S tel que

$S = x_1 + x_2 + ... + x_n$ ©

En écriture indicielle $S = \sum_{i=1}^{n} x_i$ ©

Sigma est le signe de la somme. Ceci se lit comme la somme des x_i pour i variant de 1 à n.

4) Multiplication

Pour une variable X composée de N éléments on cherche M tel que

$M = x_1 * x_2 * ... * x_n$ ©

En écriture indicielle M = $\prod_{i=1}^{n} x_i$ ©

Pi est le signe de la multiplication. Ceci se lit comme la multiplication des x_i pour i variant de 1 à n.

B. Langage statistique

Quelques définitions.

1) Population

Définition : Une population statistique est l'ensemble de référence sur lequel on effectue des observations. La population est composée d'un ensemble fini d'éléments ou d'observations (on peut tous les compter et les identifier).
Lorsque l'on démarre une étude statistique, il est très important de définir la population de façon très précise, de manière à ce que les résultats soient compréhensibles et comparables avec d'autres études menées sur le sujet. Exemple : on mène une étude sur les habitants des communes du Bas-Rhin. Est-ce que l'on prend les personnes du Chef-lieu et des hameaux où seulement celle du chef-lieu (sans les hameaux et habitations isolées). Faut-il compter les élèves en internat, les militaires…
Ainsi, en omettant de définir précisément les éléments utilisés, on peut faire mentir la statistique.
Les conclusions d'une étude statistique ne sont valables que pour la population étudiée, il ne faudra pas chercher à étendre les résultats sans les plus extrêmes précautions.

2) Individu

Définition : un individu, aussi appelé unité statistique, est un élément qui appartient à la population. Exemple : si on étudie les départements français, la France a une population de 96 départements et le Bas-Rhin est considéré comme un individu. Les populations et les individus sont de nature diverses : êtres humains (habitants, salariés, locataires…) ; objets (pièces mécaniques, voitures…) ; faits ou actes (appels téléphoniques…) ; unités de temps (jours, semaines…).

3) Echantillon

Lorsque la population est trop nombreuse ou impossible à observer dans sa totalité, on choisit un sous-ensemble de la population selon des critères qu'il faut préciser.
Définition : un échantillon est un sous-ensemble prélevé dans une population. L'analyse statistique sur échantillon est très couramment pratiquée par les instituts de sondage, ou en géomorphologie, climatologie ou en hydrologie pour des raisons évidentes de commodité, de rapidité, de coût, car on évite d'étudier toute une population.
Le problème alors c'est de savoir dans quelles conditions il est possible d'étendre les conclusions obtenues sur un échantillon à toute la population. Un échantillon non représentatif est dit « biaisé ».
Exemple : En 1936 aux Etats Unis, le républicain Lindon se présentait à l'élection présidentielle contre le démocrate Franklin Roosevelt. Deux semaines avant l'élection, le magazine Literary Digest envoya 12 millions de cartes postales pour demander aux électeurs pour qui ils allaient voter. Il reçut 2,5 millions de réponses, 57% votant pour Lindon et 43 % pour Roosevelt. On sait que la véritable élection donna le résultat opposé. Que s'est-il passé ? Le magazine a obtenu son échantillon à partir du fichier des immatriculations de

voitures et de l'annuaire du téléphone. En 1936, au coeur de la dépression, les électeurs qui possédaient une voiture et le téléphone étaient parmi les plus aisés et, en conséquence, avaient plutôt tendance à voter Républicain. De ce fait, l'échantillon n'était pas représentatif. Le grand nombre de réponses obtenu (2,5 millions) ne pouvait compenser la non-représentativité de l'échantillon.. Un échantillon représentatif de 1000 personnes, tel qu'on en utilise couramment aujourd'hui, est largement préférable pour obtenir des résultats valables. Le principe est que la taille d'un échantillon ne compensera jamais sa non-représentativité.

4) Caractère et modalité

Lorsque l'on a choisi d'étudier une population, de nombreuses questions se posent à son sujet. On va chercher à définir les principales caractéristiques de la population (ses propriétés) comme par exemple : l'âge ou la taille si on travaille en médecine ; le salaire des catégories professionnelles ; le débit des cours d'eau en hydrologie ; la température de l'air si on travaille en climatologie…
Les individus qui composent la population peuvent donc être décrits à l'aide d'un ou de plusieurs caractères.

(i) Modalité

On appelle modalité les différentes valeurs possibles du caractère.

(ii) Caractère quantitatif (variable statistique)

Le caractère est quantitatif lorsqu'on peut l'estimer en lui associant un nombre (poids, taille, salaire….).
Dans ce cas la variable peut être discrète ou continue.
La variable discrète ne prend qu'un nombre fini de modalités. Exemple : le nombre d'enfants.
La variable continue peut prendre en théorie un nombre infini de modalités. Exemple : la température, la taille. En théorie, il est toujours possible de trouver la valeur d'un élément intermédiaire entre deux tailles, deux températures…
Attention, en pratique, la plupart des caractères quantitatifs deviennent discrets en raison de la précision des appareils de mesure !
Le caractère quantitatif est toujours associé à une unité de mesure, il faudra la préciser.

(iii) Caractère qualitatif

Le caractère est dit qualitatif si on ne peut pas le mesurer (la couleur des cheveux, la profession, la marque d'une voiture).
L'ensemble des modalités que peut prendre le caractère s'appelle alors une nomenclature : l'ensemble des noms des marques de voiture forme une nomenclature, mais Renault est une modalité de cette nomenclature.
Pour être valable une nomenclature doit comprendre au minimum deux modalités !
Les rubriques de la nomenclature sont parfois codées pour une identification aisée. Exemple : A l'entrée d'un musée on demande aux gens d'indiquer leur département : ces numéros n'ont aucune signification, on ne peut pas calculer un département moyen ! Ces numéros servent à créer des catégories, des groupes. Aussi, bien que la nomenclature soit codée, il s'agit bien d'un caractère qualitatif.
La nomenclature est plus ou moins facile à mettre en place car : dans le cas des voitures on pourra prendre les marques, la couleur, tel équipement…
Par contre, on aura du mal à être exhaustif sur les catégories socioprofessionnelles, elles sont tellement variées qu'on devra les regrouper par catégories les plus homogènes possibles, d'où des difficultés : exemple : va-ton créer pour les enseignants une catégorie à eux ou va-t-on les inclure avec les cadres ?

(iv) Transformation d'une information quantitative en qualitative

Pur des évidentes, il vaut mieux se procurer une information quantitative la plus détaillée possible, mais pour des raisons pratiques, il arrive parfois qu'on veuille simplifier des formulaires si le caractère quantitatif en tant que tel ne nous intéresse pas et qu'on veuille simplement identifier des groupes : est-ce que les jeunes dans cette promotion réussissent mieux que les vieux ? On va leur demander leur âge en fixant

- deux modalités : plus de 20 ans ou moins de 20 ans
- trois modalités : –20 ans, 20-23 ans plus de 23 ans
- etc…

Ceci permettra de créer des groupes pour lesquels on calculera une moyenne aux examens.

5) Série statistique

La série statistique désigne l'ensemble des données (valeurs) obtenues pour un caractère quantitatif. Il doit y avoir autant de valeurs dans la série que d'individus étudiés (ce qui n'est pas évident à obtenir lorsque des gens choisis dans un échantillon au hasard ne répondent pas, ou ne répondent pas à certaines questions).
Il est impératif que chaque donnée puisse être rattaché à un individu clairement identifié si les caractères sont acquis séparément, ce qui permettra le regroupement des données.

Exemple : ©

1) Acquisition

Identifiant	X
Bas-Rhin	x_1
Moselle	x_2
Vosges	x_3
…	…
JJJJJJJJJ	x_n

Identifiant	Y
Vosges	y_1
Moselle	y_2
Bas-Rhin	y_3
…	…
JJJJJJJJJ	y_n

Identifiant	Z
Moselle	z_1
Bas-Rhin	z_2
Vosges	z_3
…	…
JJJJJJJJJ	z_n

2) Regroupement

Identifiant	X	Y	Z
Bas-Rhin	x_1	y_3	z_2
Moselle	x_2	y_2	z_1
Vosges	x_3	y_1	z_3
…	…	…	…
JJJJJJJJJ	x_n	y_n	z_n

3) Tableau définitif

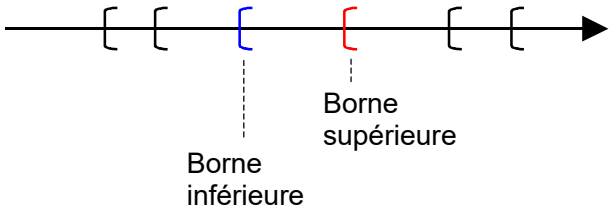
Identifiant	X	Y	Z
Bas-Rhin	x_1	y_3	z_2
Moselle	x_2	y_2	z_1
Vosges	x_3	y_1	z_3
…	…	…	…
JJJJJJJJJ	x_n	y_n	z_n

6) Classe

Quand le nombre de valeurs dans une série statistique est important il est souhaitable de les regrouper par classes.

Classe : ensemble d'unités statistiques considérées comme équivalentes placées dans une même rubrique.
Le classement consiste donc à partitionner la population. Les classes ne doivent pas avoir d'éléments en commun. Les classes ne doivent pas être trop nombreuses, sinon le classement n'apporte pas de simplification ; elles ne doivent pas être en nombre trop restreint sinon on perd trop d'information et on risque de regrouper des individus qui se ressemblent peu.
Lorsque le caractère est qualitatif, les classes dépendent de la nomenclature.
Lorsque la variable est discrète, les classes peuvent correspondre aux valeurs de la variable.
Lorsque la variable est continue, on partage la série statistique en intervalles plus ou moins réguliers.

Une variable



©

Chaque classe est définie par ses bornes inférieure et supérieure (a et b) avec (a < b) ©. En général on considère que a appartient à la classe et b n'y appartient pas. La classe s'écrit : [a ;b[©.
On parle d'intervalle fermé à gauche et ouvert à droite.

La différence entre a et b s'appelle l'amplitude de la classe et se calcule en faisant b-a©.
Le centre de chaque classe est souvent utilisés dans les calculs, comme on le verra par la suite.
Le découpage en classes est très délicat à effectuer et conditionne aussi les résultats des analyses (il est possible de tricher ou de fausser les résultats). Il existe de nombreuses méthodes de découpage en classes : classes d'amplitude constante ou variable

7) Effectif

Effectif : nombre d'éléments que contient chaque classe.
Voici une série statistique discrète qui comprend 6 observations, celle de la variable « Notes » ©

Notes	5	7	8	5	11	8

On va d'abord classer la variable note : comme la variable est discrète on va créer 4 classes de notes.
On va ranger les classes par ordre croissant, comme on a fait un regroupement, il va falloir indiquer le nombre d'individus que contient chaque classe.

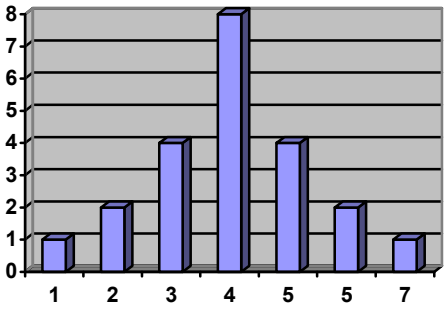
Notes	5	7	8	11
Effectif	2	1	2	1
Effectif cumulé	2	3	5	6

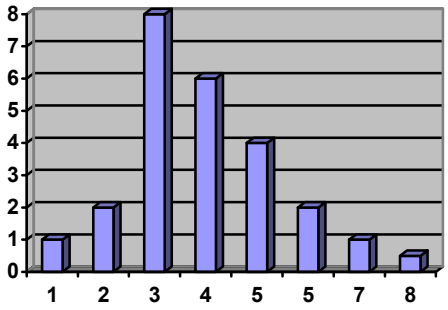
La ligne des effectifs détermine une distribution statistique d'effectif total (N) = somme de tous les effectifs (ici N = 6) et de k classes.
L'intérêt d'avoir rangé les valeurs par ordre croissant c'est que si on se pose la question : quel est le nombre d'élèves ayant eu une note inférieure ou égale à 7, on compte le nombre d'élèves ayant obtenu 5 + le nombre d'élèves ayant obtenu 7.
Si la variable est classée dans le désordre la réponse peut être longue à trouver, mais si elle est classée dans l'ordre croissant, c'est simple.

D'autant plus qu'il est possible de travailler sur les effectifs cumulés. Le cumul des effectifs permet de connaître instantanément le nombre d'individus ayant une valeur inférieure ou supérieure à une certaine valeur.
Attention cependant, car le découpage des classes détermine la répartition des effectifs (Voir exemple des notes obtenues par des candidats et le découpage en classes de 1,2,4,5 points) ! ©

8) Distributions simples

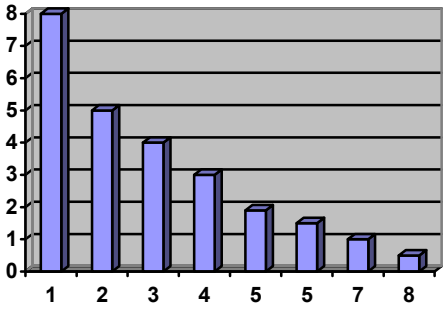
Les distributions symétriques décroissent symétriquement de part et d'autre d'un maximum central. Ces distributions sont assez rares. Les paramètres de tendance centrale (comme la moyenne) ont une signification.

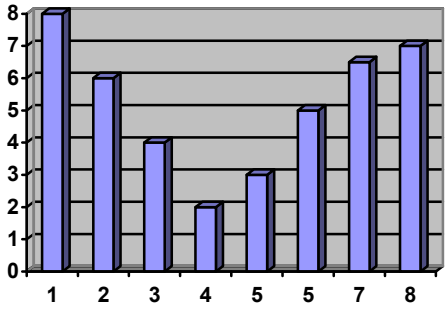




©

Les distributions asymétriques à un seul sommet : les effectifs décroissent plus rapidement d'un côté que de l'autre du maximum. Se sont les plus fréquentes, on a assez fréquemment un étalement vers les valeurs les plus élevées de la variable. Certains paramètres de tendance centrale sont significatifs.
Les distributions hyperboliques dans lesquelles la fréquence maximum se présente à une extrémité.
Les distributions en U





©

Montrer des exemples de distribution réels©

9) Fréquence

Nombre par lequel on désigne l'importance relative d'un événement ou d'une observation, alors que l'effectif donne l'importance absolue d'un évènement.
Si on dispose de (N) observations et que l'observation particulière (i) se produit n fois, sa fréquence est : $f_i = n_i / N$.

Notes	5	7	8	11
Effectif	2	1	2	1
Effectif cumulé	2	3	5	6
Fréquence	0.33	0.17	0.33	0.17
Fréquence cumulée	0.33	0.5	0.83	1

La fréquence doit être comprise entre 0 et 1 : $0 \leq f_i \leq 1$. ©
La somme de toutes les fréquences doit être égale à 1.

$$\sum_{i=1}^{k} f_i = 1$$ ©

Les fréquences sont souvent exprimées en pourcentage.
Comme je le disais, la fréquence indique l'importance relative de chaque classe alors que les effectifs nous donnent l'importance absolue. Selon le cas on peut utiliser l'une ou l'autre des formulations.
On peut toujours calculer des fréquences à partir des effectifs, mais l'inverse n'est vrai que si on connaît l'effectif total (N) : $n_i = f_i * N$. ©
Il est possible de masquer un nombre insuffisant d'observations en utilisant les fréquences.
La fréquence cumulée indique l'importance relative des observations inférieures ou égales à un certain seuil.

10) Erreur statistique

On peut distinguer deux types d'erreurs

(i) Les erreurs aléatoires

Qui surviennent par exemple quand vous faites un sondage à partir d'un échantillon et non pas à partir d'une population, si l'échantillon est tiré au hasard, les résultats obtenus oscillent autour de la valeur vraie. Les erreurs aléatoires surviennent aussi au moment de la collecte des données, normalement elles se compensent.

(ii) Les erreurs systématiques

C'est le cas quand les sondés ne répondent pas sincèrement aux questionnaires, ou lorsque l'enquêteur n'est pas neutre… ! Dans ce cas les résultats sont biaisés.

III. ANALYSE ELEMENTAIRE DES SERIES STATISTIQUES

Lorsque une série comporte un grand nombre de valeurs, on cherche à la résumer à l'aide de quelques nombres significatifs appelés paramètres. Ces paramètres permettront aussi de comparer plusieurs séries statistiques entre elles : Les températures de Strasbourg par rapport à celles de Lyon par exemple.
On a constaté précédemment que les données d'une série statistique accusent une certaine accumulation des effectifs autour d'une valeur particulière du caractère et présentent un étalement plus ou moins grand des effectifs de part et d'autre de cette valeur.
Ainsi la description d'une série statistique doit être faite sous trois angles :

- analyse du paramètre central (position) de la série ;
- dispersion ou fluctuation des observations autour de cette valeur centrale ;
- forme (symétrie, aplatissement) de la distribution de la série.
- dans certaines conditions on peut calculer un indice de concentration

A. Paramètres de position d'une série

L'idée est de caractériser une série par un nombre unique, représentatif de la série, de telle sorte que, la comparaison de deux séries se ramène à la comparaison de deux nombres.

1) Moyenne

La moyenne d'une série X se note conventionnellement $\bar{x}$ et se lit x barre ou x moyenne.

(i) Données brutes

Une série de données brutes©

Notes (X)	5	7	8	5	11	8

On applique la formule de la moyenne arithmétique

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{N} \qquad = \frac{1}{N}(x_1 + x_2 + ... + x_n)$$ ©

D'où $$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$$ ©

$\bar{x}$= 7.3

(ii) Série classée

Lorsque la série est classée il est inutile de l'étendre pour obtenir la moyenne. ©

Notes (X)	5	7	8	11
Effectif	2	1	2	1

Variable = X
Effectif total = N
Nombre de classes = K

Effectif par classe = E

On calcule une moyenne pondérée : chaque note est pondérée par son effectif. Le poids de chaque valeur dépend de son effectif

$$\bar{x} = \frac{e_1 x_1 + e_2 x_2 + ... + e_k x_k}{N} \qquad \frac{1}{N}\left(e_1 x_1 + e_2 x_2 + ... + e_k x_k\right)©$$

D'où $$\bar{x} = \frac{1}{N}\sum_{i=1}^{k}\left(e_i x_i\right)©$$

$\bar{x}$ = (2*5 + 1*7 + 2*8 + 1*11) / 6 = 7.3©

Important ! Si on a des classes avec des bornes, on remplace xi est par le centre de la classe, mais alors on obtient une approximation de la moyenne.

(iii) Moyenne à partir de classes dont on connaît les moyennes

Si on a une population de N individus divisée en p classes pour lesquelles on a une moyenne (m) pour lesquelles on connaît les effectifs (e), alors la moyenne générale se calcule selon la méthode pondérée :

©

X	C_1	C_2	C_3	…	C_p
Moyenne M	m_1	m_2	m_3		m_p
Effectif E	e_1	e_2	e_3		e_p

Variable = X
Effectif total = N
Nombre de classes = K
Effectif par classe = E
Moyenne par classe = M

$$\bar{x} = \frac{e_1}{N} m_1 + \frac{e_2}{N} m_2 + ... + \frac{e_k}{N} m_k \qquad = \frac{1}{N}\left(e_1 m_1 + e_2 m_2 + ... + e_k m_k\right)©$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k}\left(e_i m_i\right)©$$

Même formule que précédemment, mais on remplace le centre de la classe par la moyenne de la classe. Cette méthode est meilleure que la précédente quand on a des classes avec des bornes.

(iv) Moyenne à partir de fréquences :

On voit ici que $\frac{e_k}{N}$ est une fréquence. Donc si on a des classes et des fréquences il est possible de calculer la moyenne de la façon suivante : pour une variable discrète par exemple

Classes de X	Fréquences
x_1	f_1
x_2	f_2
…	…
x_n	f_n

$\bar{x} = \sum_{i=1}^{k} (f_i x_i)$ ©

En remplaçant les x par le centre des classes si on a des bornes ou mieux, par la moyenne de la classe.

(v) Remarques

Quelques propriétés mathématiques de la moyenne :
Si X et Y sont deux séries, la moyenne des nombres $x_1 + y_1$, x_2+y_2,…., x_n+y_n est égale à : $\bar{x} + \bar{y}$.©
Si a est une constante quelconque et X une série de nombres, la moyenne des nombres $x_1 + a$, $x_2 + a$,…, $x_n + a$ est égale à : $\bar{x} + a$. ©
Si a est une constante quelconque et X une série de nombres, la moyenne des nombres a x_1, a x_2,…,a x_n est égale à : a $\bar{x}$.©

$\sum_{i=1}^{n} (x_i - \bar{x})^2$ minimale. ©

Ce qui a pour conséquence que la moyenne est le centre de gravité d'une distribution (illustration par le graphique).
La moyenne est sensible aux valeurs extrêmes qui ont tendance à attirer le centre de gravité vers elles. Or ces valeurs extrêmes sont souvent peu significatives, exceptionnelles, voire aberrantes, c'est pourquoi parfois on les élimine carrément des séries, ou alors on calcule une valeur moyenne partielle sans les valeurs extrêmes, mais qui sont conservées pour d'autres traitements.

2) Médiane

La médiane désigne le nombre qui permet de couper la population étudiée en deux groupes contenant le même nombre d'individus. La médiane renseigne sur la répartition des effectifs : 50% des individus ont une valeur inférieure à la médiane et 50% une valeur supérieure. La médiane à l'avantage de ne pas dépendre des valeurs extrêmes (si on élimine trois valeurs quelconques d'un côté et 3 valeurs extrêmes de l'autre, la médiane ne change pas, alors que la moyenne sera complètement différente.

(i) Calcul de la médiane d'une série brute :

1) Ranger la série de nombres par ordre croissant, en répétant les nombres©

X Non trié	1	2	5	1	8	9

X Trié	1	1	2	5	8	9

2) si l'effectif (N) de la population est impair, trouver le nombre x_i situé au milieu de la suite, tel que son indice i = (N+1)/2 ; Ici i = 3 et la valeur de la médiane est 2. ©

3) si l'effectif (N) de la population est pair, la médiane est le nombre situé au centre de l'intervalle formé par les valeurs de rang i = $\frac{N}{2};\frac{N}{2}+1$, cet intervalle s'écrit $\left[x_{\frac{N}{2}};x_{\frac{N}{2}+1}\right]$ ©

Il faut donc trouver le milieu de [x_3 ;x_4] soit le milieu de [2;5] donc M = 3.5.
La médiane n'est donc pas nécessairement une valeur de la suite .
Il peut arriver que l'intervalle médian soit du type [a ;a], dans ce cas M vaut a.

Cette méthode implique que pour trouver la médiane d'une série classée il faut étendre la série. Exemple : ©

X	50	45	30	60	61
Effectif	2	3	2	2	1

D'abord on range les valeurs dans l'ordre croissant :

30, 30, 45, 45, 45, 50, 50, 60, 60, 61. ©
Comme N est pair (N = 10), la médiane est le milieu de l'intervalle médian délimité par la cinquième et sixième valeur [x_5 ;x_6] soit [45 ;50], donc M = 47.5. ©
Cette méthode est la méthode idéale et la meilleure, mais longue et fastidieuse si on a beaucoup de données, en plus il faut posséder les données originales.

(ii) Série classée

Que l'on se rassure, pour une série classée, inutile d'étendre la série !
Si les données sont regroupées par classes, pour trouver la médiane, il faut recourir à un procédé d'interpolation linéaire.
Exemple : on effectue des essais sur un échantillon de 199 ampoules (N=199) pour tester leur durée de fonctionnement. Les résultats sont regroupés en classes d'amplitude 100 heures. ©

Classe X (h)	[1200;1300[	[1300;1400[	[1400;1500[	[1500;1600[	[1600;1700[
Effectif	30	50	70	30	19
Effectif cumulé	30	80	150	180	199

Comme N est impair, la médiane correspond à la valeur x_{100}. Cette valeur appartient à la classe [1400;1500[. La valeur de la médiane se situe donc entre 1400 et 1500 heures.
On va utiliser une interpolation linéaire(équation de proportionnalité) :
La formule s'écrit donc $M = V + \frac{D}{E}A$ ©
ainsi dans l'interpolation la valeur cherchée (M) dépend de la valeur la plus faible de la classe (V), de la différence entre le rang de la valeur recherchée et les effectifs cumulés des classes précédentes (D), de l'effectif de la classe (E) et de l'amplitude de la classe (A). ©
d'où $M = 1400 + \frac{20}{70}100$ = 1428.57 h©
Inconvénient majeur : dans ce cas, le résultat est une approximation de la vraie valeur.
A part son pouvoir descriptif, la médiane est peu employée dans les calculs statistiques.
Elle possède la propriété mathématique suivante :

$\sum_{i=1}^{n}|x_i - a|$ minimale si a est la médiane. ©

3) Mode, classe modale

Dans de nombreux cas on veut connaître la valeur la plus fréquente de la variable : dans une population, parmi toutes les tailles, quelle est celle qui revient le plus souvent ? Dans une station de traitement des eaux, on teste toutes les semaines la teneur en nitrate de l'eau, quelle est la valeur qui revient le plus souvent ?
On appelle mode d'une série statistique une valeur ou la classe dont l'effectif ou la fréquence est le plus grand. Le mode est donc la valeur la plus probable. Il peut y avoir plusieurs modes dans une série, comme dans le cas des séries bi modales. Ces séries feront l'objet d'un traitement spécial.
Lorsque la série statistique se présente sous forme de classes, la classe qui présente le plus grand effectif est appelée classe modale.

Voici une série statistique discrète des vitesses de véhicules ©

Vitesses (Km/h)	70	72	74	75	78	80	83
Effectif	2	1	2	2	1	3	1

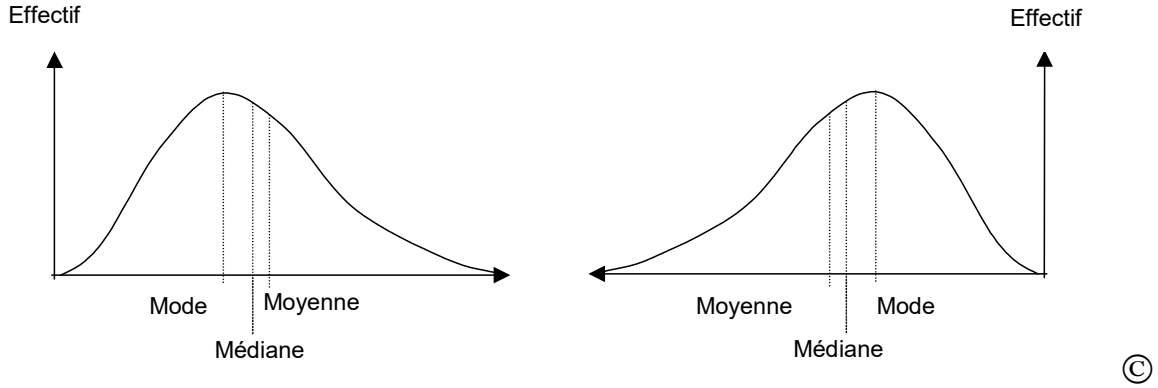
Son mode est 80, mais on va la classer par classes d'amplitude 3

Classe de vitesses (Km/h)	[70 ;73[	[73 ;76[	[76 ;79[	[79 ;82[	[82 ;85[
Effectif	3	4	1	3	1

La classe [73 ;76[est la classe modale de cette série classée. La classe modale ne correspond pas au mode. La classe modale est sensible à la façon dont on procède au classement.

4) Comparaison des paramètres de position

Dans les distributions symétriques, le mode, la moyenne et la médiane ont la même valeur. Lorsque la distribution est légèrement asymétrique, et unimodale, la médiane se trouve entre la moyenne et le mode.



©

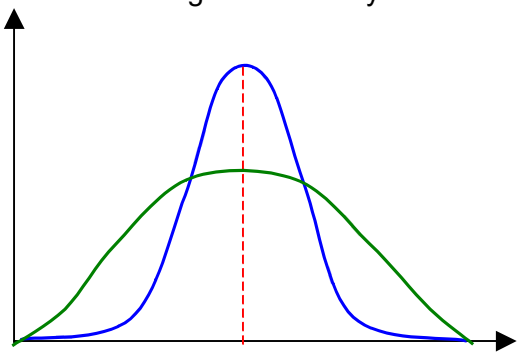
Pourquoi ?
On a vu que la moyenne est très influencée par les valeurs extrêmes et le centre de gravité se déplace du côté où elles sont les plus nombreuses, donc à droite.
Puisque le distribution est asymétrique, il y a plus de valeurs à droite du mode qu'à gauche, donc la médiane est décalée à droite, mais sa valeur est inférieure à celle de la moyenne.

En étudiant de nombreuses distributions, on s'aperçoit que la médiane ne se positionne pas n'importe où : sa « distance » au mode est environ double de sa distance à la moyenne, donc :

Mode = 3*Médiane – 2 * Moyenne©
Cette règle est intéressante car elle permet d'évaluer statistiquement et approximativement la valeur du mode en connaissant les 2 autres.
Mais attention, cette règle n'est pas toujours vraie, elle n'est valable que pour les distributions légèrement asymétriques.

B. Les paramètres de dispersion

Lorsque les valeurs de la série ne sont pas trop différentes, la moyenne permet d'avoir une idée assez juste de la série. Mais si certaines valeurs sont très différentes, la distribution devient hétérogène et la moyenne ne donne pas une idée juste de la réalité.



Faible dispersion	Forte dispersion

©

De ce fait la moyenne arithmétique n'a pas toujours de signification concrète et doit toujours être accompagnée d'une caractéristique de dispersion.

1) Dispersion des effectifs : les quantiles

La médiane donne une idée de la répartition des effectifs (moitié/moitié). Mais parfois on souhaite avoir des renseignements plus précis sur la répartition des effectifs. Par exemple, le salaire brut médian versé en France est de 20389 Euros/an (une moitié gagne moins et l'autre gagne plus).
L'idée est de partager les effectifs en groupes de même effectif chacun : les quartiles (4 groupes de 25%), les quintiles (5 groupes de 20%) et les déciles (10 groupes de 10%)©.
Les quartiles désignent les nombres qui permettent de couper la population étudiée en 4 groupes contenant le même nombre d'éléments (25% des effectifs chacun).

(i) Cas des données brutes

Voici une série de 23 nombres (n = 23) rangés dans l'ordre croissant©.

```
4  4  4  4  4  4  7  7  7  7  7 10 10 10 10 10 10 10 13 13 13 13 16
   25%      |    25%       |        25%       |        25%
            Q1              M                  Q3
```

1) on divise la série selon la médiane (12ème élément)
2) on divise les deux demi séries en deux effectifs égaux ou selon la plus petite valeur qui convient

La valeur de Q_1 est la valeur du premier quartile : au moins 25% des effectifs ont une valeur du caractère inférieure ou égale à 4
La valeur de Q_3 est la valeur du troisième quartile : au moins 75% des effectifs ont une valeur du caractère inférieure ou égale à 10.

Comment trouver facilement les quartiles ? ©
Si i = N/4 est un entier, alors Q_1 est le terme qui occupe le rang i et Q_3 est le terme qui occupe le rang 3i.
Si i = N/4 n'est pas un entier, Q_1 est le terme de rang immédiatement supérieur à i (ici N/4 = 5.75 donc le 6ème rang) et Q_3 est le terme de rang immédiatement supérieur à 3N/4 (ici 3N/4 = 17.25 donc le 18ème rang).

(ii) Cas d'une série classée, après calcul des fréquences

Classe X (h)	[0 ;2[	[2 ;4[	[4 ;6[	[6 ;8[	[8 ;10[	[10 ;12[	[12 ;14[
Fréquence	0.12	0.15	0.18	0.24	0.14	0.10	0.07
Fréquence cumulée	0.12	0.27	0.45	0.69	0.83	0.93	1

©
Q_1 doit se trouver dans la classe où se trouve la fréquence cumulée 0.25, donc dans la classe [2 ;4[on va trouver Q_1 par interpolation linéaire :

$$Q_1 = 2 + \frac{0.13}{0.15}2 = 3.73$$ ©

On procède de même pour Q_3 et pour la médiane.

$$Me = 6 + \frac{0.05}{0.24}2 = 6.41$$ ©

$$Q_3 = 8 + \frac{0.06}{0.14}2 = 8.86$$ ©

On voit ici qu'entre la médiane et Q1 les effectifs sont répartis sur un plus grand intervalle (2.68) qu'entre la médiane et Q3 (2.45)

2) Dispersion des valeurs :

(i) Etendue

L'étendue est la différence entre les valeurs extrêmes du caractère étudié :

$E = x_{max} - x_{min}$ ©
Plus l'étendue est faible, plus le caractère étudié est homogène. Exemple, à Abidjan l'amplitude thermique annuelle à partir des données moyennes mensuelles est de 3.8° (27.8-24) alors qu'à Leningrad elle est de 26.3° (18.4-(-7.9)). Dans un cas le climat est plus contrasté que dans l'autre.
Méfiance, car l'étendue traduit la dispersion de la série, mais à l'aide de deux valeurs uniquement, et en plus de manière imparfaite puisque les valeurs extrêmes sont souvent peu significatives, car exceptionnelles, accidentelles, voire fausses.
Autre inconvénient, c'est que pour une même étendue, les valeurs peuvent être plutôt regroupées autour de la moyenne ou bien réparties dans l'étendue. Il faut donc traduire cette dispersion.

(ii) L'intervalle interquantile

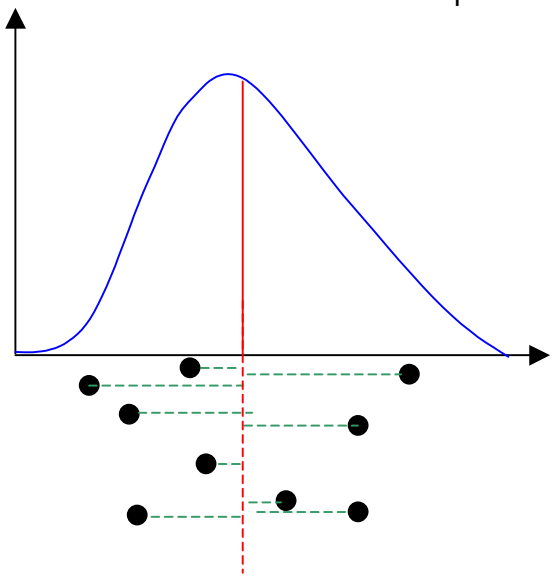
On va se baser sur les quantiles pour mesurer la dispersion au centre de la distribution sans tenir compte des extrêmes.

On pourra alors utiliser l'intervalle $[Q_1 ; Q_3]$ ©. La valeur de cet intervalle est appelée interquartile ($I = Q_3 - Q_1$) ©. Plus I est grand, plus la dispersion est élevée et le phénomène hétérogène. Cependant il ne tient compte que de 50% des effectifs, ce qui entraîne une perte de l'information.

L'intervalle $[D_1 ; D_9]$ aussi appelé interdécile ($I = D_9 - D_1$) © est basé lui sur 80% des valeurs et a l'intérêt de n'éliminer que les vraies valeurs extrêmes.

(iii) L'écart absolu moyen

En fait ce que l'on cherche à faire s'est mesurer la dispersion des valeurs (x_i) de la série statistique autour de la position centrale de la moyenne ($\bar{x}$). Donc on veut une quantité qui mesure l'écart moyen par rapport à la valeur moyenne. Pour cela, la logique voudrait que l'on mesure la différence entre chaque valeur x_i et la moyenne, puis on fait la somme de toutes ces différences et on divise par le nombre de différence pour obtenir un écart moyen.



©

Par calcul $e_m = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)$ Or le problème c'est que $\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$.©

Car étant donné les caractéristiques de la moyenne, par définition, les différences positives compensent les différences négatives dans la distribution

Pour éviter cela, certains utilisent alors la valeur absolue de la différence et font :

$$e_m = \frac{1}{N}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| ©$$

On obtient ainsi l'écart absolu moyen. Cet écart absolu moyen est peu utilisé en statistique, car la valeur absolue se prête mal aux calculs compliqués, on lui préfère une autre formulation.

(iv) La variance

En effet on préfère utiliser le carré de la différence $\left(x_i - \bar{x}\right)^2$ © dont on sait en plus que la somme est non nulle, mais minimale !

En calculant la moyenne de ces écarts on obtient la variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ ©

Attention, cette formule de variance n'est valable que si les données concernent une population dans son ensemble.
Si vos données correspondent à un échantillon, tiré d'une population

La variance s'écrit $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ ©

Pourquoi diviser par N-1 ?
Supposons que nous ne disposions que d'un échantillon de 1 individu.
On peut estimer le poids moyen de la population : ce sera le poids de l'individu (ex : 65 kg).
L'écart-type calculé avec N donnerait une dispersion nulle, ce qui suggère que toute la population pèse précisément 65 kg !
Donc il faut utiliser N-1 pour indiquer que nous ne pouvons pas estimer la dispersion dans la population si notre échantillon ne comporte pas au moins 2 individus, (car on ne peut pas diviser par zéro).
On constate que si le N de l'échantillon devient grand (> plusieurs milliers), alors

$\frac{1}{N-1} \approx \frac{1}{N}$ ©

Pour faciliter les calculs : $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2$ ©

Exemple à partir des données brutes : ©

Rang	X	X^2
1	50	2500
2	60	3600
…	…	…
n	x_n	x_n^2
moyenne	$\bar{x}$	$\frac{1}{N}\sum_{i=1}^{n} x_i^{\,2}$

Si on dispose d'une variable classée dont on connaît les fréquences (f) associées à chaque classe :

$\sigma^2 = \sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2$ ©

Exemple : ©

Rang	Classe X	Fréquence	FX	FX^2
1	[0 ;20[	0.1	1.0	10
2	[20 ;40[	0.05	1.5	45
…		…	…	
k	$[a_k ; b_k[$	f_k	$f_k(a_k+b_k)/2$	$f_k((a_k+b_k)/2)^2$
moyenne			$\sum_{i=1}^{k}\left(f_i x_i\right)$	$\sum_{i=1}^{k} f_i x_i^{\,2}$

La variance n'est qu'une étape intermédiaire de calcul, elle n'a pas de signification, mais elle est impliquée dans beaucoup de calculs qui cherchent à établir des relations entre variables (voir par la suite)

(v) L'écart-type

La valeur de l'écart-type se déduit de celle de la variance. Il s'écrit σ, et il se calcule simplement comme la racine carrée de la variance.
Pour une population

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad ©$$

Pour un échantillon

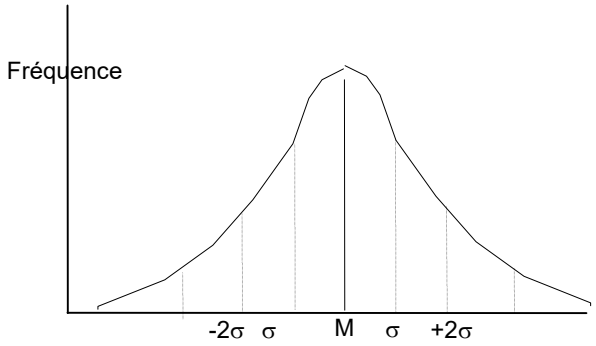
$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad ©$$

L'écart-type est une valeur très importante car il nous donne une idée de la dispersion d'une variable autour de sa moyenne arithmétique. C'est aussi une valeur qui intervient dans de nombreux calculs.
D'après les études menées, si la distribution de la population n'est pas trop asymétrique, on sait que dans l'intervalle© :

] $\bar{x}$ - σ ; $\bar{x}$ + σ [nous trouvons 68.3 % des effectifs

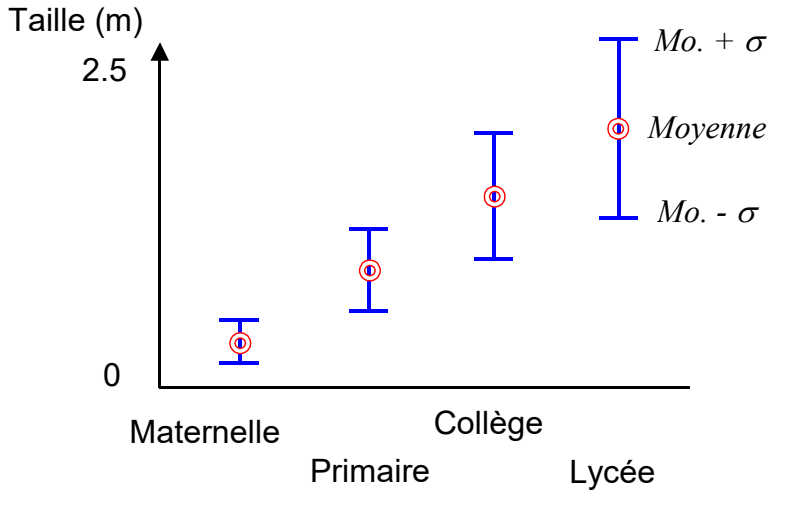
] $\bar{x}$ - 2σ ; $\bar{x}$ + 2σ [nous trouvons 95.4 % des effectifs

] $\bar{x}$ - 3σ ; $\bar{x}$ + 3σ [nous trouvons 99.7 % des effectifs



3) Coefficients de dispersion relative

Les caractéristiques de dispersion que nous avons utilisées jusqu'à maintenant sont exprimés dans les mêmes unités que la variable étudiée. Si la variable est une longueur, l'écart-type est une longueur, si elle est en Euros, l'écart-type est en Euros…
Ceci est un inconvénient quand on veut comparer la dispersion de caractères différents, ou encore de séries d'un même caractère exprimées dans des unités différentes (la dispersion de la taille des anglais et français mesurée en pouce ou en centimètres).
En outre les variables peuvent être telles que le niveau moyen de l'une des séries est nettement différent du niveau moyen de l'autre série (salaire moyen des ouvriers et des cadres par exemple), ce qui influence évidemment les valeurs de l'écart-type. ©



©

C'est pourquoi il faut éliminer l'influence de la variable étudiée, son unité de mesure et son niveau général. On utilise alors un coefficient de dispersion relatif.

(i) Le coefficient interquartile ou interdécile

$\frac{Q_3 - Q_1}{Q_2}$ © c'est un nombre abstrait, CAD indépendant de l'unité de la série. Pour le coefficient interdécile, la logique est la même $\frac{D_9 - D_1}{D_5}$ ©

(ii) Coefficient de variation

Le plus souvent on compare la valeur obtenue par l'écart-type avec la valeur moyenne pour obtenir ce que l'on appelle le coefficient de variation : C tel que :

$C = \frac{\sigma}{\bar{x}}$ ©

On l'exprime en %. Un coefficient de variation de 0,21 signifie que l'écart-type vaut 21% de la moyenne arithmétique. L'écart-type nous donne une mesure absolue de la dispersion, alors que le coefficient de variation nous donne une mesure relative, par rapport à la valeur moyenne.

(iii) Exemple

Exemple : précipitations annuelles (mm) entre 1941-1970©

	Moyenne	Ecart-type	C (%)	Coefficient interquartile
Bilma (Niger, climat saharien)	19	15	79	1.2
Niamey (Niger, climat sénégalien)	592	124	21	0.2
Kumasi (Ghana, climat soudanien)	1511	281	19	0.3

La dispersion absolue (écart-type) est plus grande à Kumasi qu'à Niamey, mais la dispersion relative est du même ordre : donc la variabilité absolue des phénomènes pluvieux est

proportionelle à la valeur centrale. Par contre à Bilma, on obtient des valeurs élevées, alors que la dispersion absolue est faible.
Attention, cependant, ces coefficients deviennent de moins en moins valables quand les valeurs de la moyenne se rapprochent de 0 ! ! ! ! !

4) Entropie relative

En théorie de l'information, l'entropie mesure le degré d'incertitude sur la réalisation d'un événement parmi d'autres. Si tous les évènements ont la même probabilité de se réaliser, l'incertitude est la plus grande et on dit que l'entropie est maximale.
Mais si un événement a une très forte probabilité de se produire par rapport aux autres, l'incertitude diminue, on sait que c'est cet événement qui va arriver, l'entropie est minimale.
Ainsi en utilisant des fréquences statistiques au lieu de probabilités on peut utiliser l'entropie comme mesure de dispersion. Bien sûr ceci ne peut se calculer qu'à partir de variables ayant des classes et pour lesquelles on a calculé des fréquences.

$$E = \frac{-\sum_{i=1}^{k} f_i . log(f_i)}{log(k)} \text{ ©}$$

k = nombre de classes
f = fréquence de la classe i
log = logarithme de base 10 et comme le log(0.3) est négatif, on transforme le résultat en valeur positive par le signe -

Comme résultat, E varie entre 0 et 1. Quand E vaut 0 les données sont concentrées sur une seule valeur et quand E vaut 1 elles sont toutes dispersées avec les mêmes fréquences.

Exemple : ©
Evolution de la population départementale en région Rhône-Alpes :

	1861		1946		1975	
	Habitants	*f*	Habitants	*f*	Habitants	*f*
Ain	370	0.109	307	0.089	376	0.080
Ardèche	389	0.114	255	0.074	256	0.054
Drôme	327	0.096	268	0.077	362	0.076
Isère	578	0.170	574	0.166	860	0.182
Loire	537	0.158	632	0.183	696	0.147
Rhône	662	0.194	919	0.265	1430	0.302
Savoie	275	0.081	236	0.068	305	0.064
Haute-Savoie	267	0.078	271	0.078	448	0.095
Total	3405	1	3462	1	4733	1
Ecart-type	138.5		233.4		270.9	
C	0.49		0.81		0.69	
Entropie	0.98		0.94		0.92	

L'écart-type augmente et donne l'impression que la dispersion augmente. Or ceci est un pur effet mécanique, puisque la population totale augmente. Quand les effectifs augmentent, il est normal que l'écart-type augmente, même si les proportions par département restent inchangées.
Le coefficient de variation nous apprend que dispersion par rapport à la valeur moyenne est plus faible en 1861
L'entropie, qui décroît, montre une tendance au regroupement de la population dans certains départements, et qu'il y a une modification des proportions (des fréquences). (voir graphique) ©

5) Série centrée et réduite (standardisation)

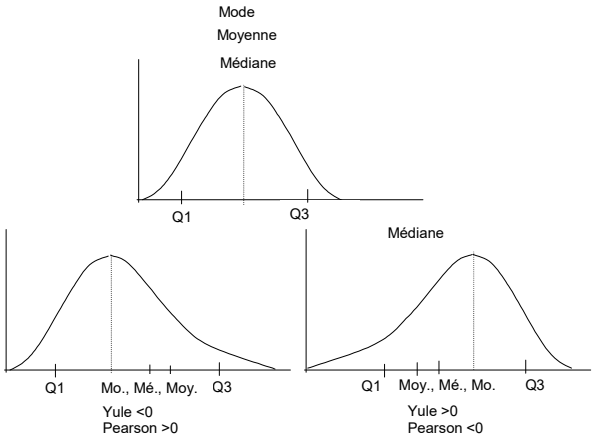
Une série est dite centrée et réduite lorsque sa moyenne vaut 0 et son écart-type vaut 1. Si X est une série statistique de moyenne $\bar{x}$ et d'écart-type σ alors la série X' est centrée et réduite si $X'_i = (x_i - \bar{x}) / \sigma$.©
Donc pour transformer une série (X) en série centrée et réduite (X') il suffit de retrancher la moyenne et de diviser par l'écart-type. On pourra ensuite vérifier que $\bar{x'}$ vaut 0 et $\sigma_{x'}$ vaut 1©.
L'objectif : il est plus simple de travailler sur des séries centrées et réduites : les calculs se simplifient. On peut facilement comparer des séries entre elles puisque elles ont la même moyenne et le même écart-type.

C. Paramètres de forme

Nous avons vu qu'une distribution est dite symétrique si les observations sont également dispersées de part et d'autre de la valeur centrale. Dans le cas contraire, la distribution est dite asymétrique ou dissymétrique.
La distribution peut également être plus ou moins aplatie au niveau de la valeur centrale en fonction de la concentration des effectifs.
Il faut caractériser la symétrie et l'aplatissement d'une distribution au moyen de nombres indépendants des unités de mesures.

1) La dissymétrie

Dans une distribution symétrique, les valeurs de mode, médiane moyenne sont confondues et les quartiles sont équidistants de la valeur centrale. Ce n'est pas le cas dans les distributions dissymétriques.



©

Aparté 09. **Indices de forme**

1)

- **Coefficient d'asymétrie**

On pose $\beta_1 = \dfrac{\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^3}{[\sigma(X)]^3}$.

On a $\beta_1 > 0$ si la distribution est étirée à droite, $\beta_1 < 0$ si la distribution est étirée à gauche, $\beta_1 = 0$ si la distribution est symétrique (par rapport à la moyenne).

- **Coefficient d'aplatissement (appelé aussi kurtosis)**

On pose $\beta_2 = \dfrac{\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^4}{3[\sigma(X)]^4}$

Le coefficient 3 au dénominateur est introduit pour que les distributions de probabilités normales (ou gaussiennes) aient un coefficient d'aplatissement égal à 1.

On a donc $\beta_2 = 1$ lorsque la distribution est de même aplatissement qu'une distribution normale, $\beta_2 < 1$ lorsque la distribution est plus plate, $\beta_2 > 1$ lorsque la distribution est moins plate.

Ce coefficient est relativement difficile à interpréter car visuellement on a l'impression que, pour une même moyenne, la distribution est d'autant plus plate que l'écart type est grand alors

(i) Le coefficient de Pearson

On va utiliser la position de la moyenne par rapport au mode pour caractériser l'asymétrie : on voit que dans une distribution dissymétrique la moyenne se déplace par rapport au mode, à droite ou à gauche, car elle est attirée par le grand nombre de valeurs extrêmes. Alors on va calculer la différence entre la moyenne et le mode et rapporter ceci à l'écart-type pour obtenir un chiffre abstrait. Dans ce cas le coefficient d'asymétrie (s) s'écrit :

$s = \dfrac{\overline{x} - Mo}{\sigma}$ ©

Comme la détermination du mode est malaisée par le calcul, on sait que :

$Mo = 3Me - 2\overline{x}$ d'où $s = \dfrac{3(\overline{x} - Me)}{\sigma}$ ©

Ce coefficient n'est valable que pour les distributions modérément asymétriques. Il peut varier de –3 à +3 et vaut 0 pour une courbe symétrique. Le signe de ce coefficient nous donne l'asymétrie et sa valeur nous renseigne sur l'ampleur de la déformation :
Lorsqu'il est inférieur à 0 la courbe est étalée à gauche (la valeur de la moyenne est inférieure à celle du mode).
Lorsqu'il est supérieur à 0, la courbe est étalée à droite (la valeur de la moyenne est supérieure à celle du mode..

(ii) Le coefficient de Yule

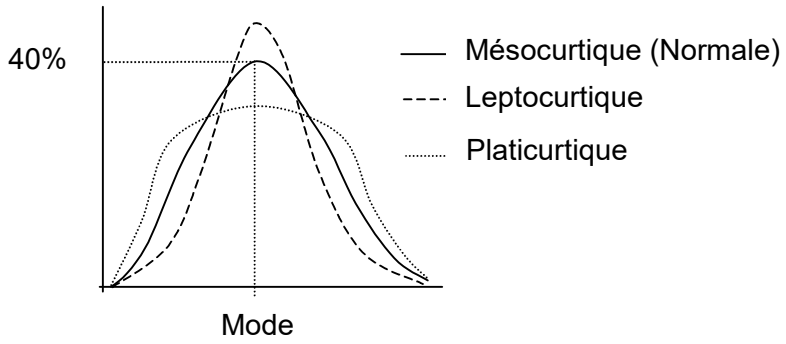
Il consiste à comparer l'étalement à gauche de la distribution par rapport à l'étalement à droite en se basant sur la position de la médiane dans l'intervalle interquartile.
Nous avons vu que la médiane est attirée du côté où se trouvent les grands effectifs.
Lorsque la distribution n'est pas symétrique, la médiane ne se situe pas au milieu de l'intervalle interquartile. On va donc mesurer l'écart entre la médiane et les bornes de cet intervalle:

L'étalement à gauche est mesuré par ($Me-Q_1$) et l'étalement à droite par (Q_3-Me). Ensuite on fait la différence de ces étalements et on divise par leur somme :

$$s = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$ Finalement $$s = \frac{Q_1 + Q_3 - 2Me}{Q_3 - Q_1}$$ ©

S varie de –1 à +1 et un coefficient de 0 traduit une symétrie parfaite.
Lorsqu'il est inférieur à 0, la distribution est étalée à droite.
Lorsqu'il est supérieur à 0 la distribution est étalée à gauche.

2) L'aplatissement



©

D. Paramètres de concentration

Cette notion tient une place importante dans les études économiques : on parle de concentration des entreprises, concentration de la richesse…
Pour déterminer la concentration il faut suivre plusieurs étapes :
1) Trouver la médiane des valeurs.
2) Trouver la médiale
3) Calculer l'indice de concentration

$$IC = \frac{|médiale - médiane|}{Etendue}$$ ©

Exemple :

Classes de salaires horaires (Euros)	Effectifs	Effectifs cumulés	Masse salariale : effectif * salaire	Masse salariale cumulée
[2-4[	5	5	15	15
[4-6[	8	13	40	55
[6-8[	12	25	84	139
[8-10[	10	35	90	229
[10-12[	8	43	88	317
Total	43		317	

1) Trouver la médiane des valeurs.
L'effectif total est de 43. La médiane correspond donc à la (43+1)/2 = 22ème valeur. Cette valeur est comprise dans la classe [6-8[d'après la colonne des effectifs cumulés.

La formule s'écrit donc (équation de proportionnalité) $M = V + \frac{D}{E} A$

ainsi dans l'interpolation la valeur cherchée (M) dépend de la valeur la plus faible de la classe (V=6), de la différence entre le rang de la valeur recherchée et les effectifs cumulés des classes précédentes (D=9), de l'effectif de la classe (E=12) et de l'amplitude de la classe (A=2). La médiane vaut 7.5.

2) Trouver la médiale des valeurs du caractère étudié. La médiale est telle que la valeur de tous les caractères supérieurs constitue une moitié et la valeur des caractères inférieurs une autre moitié. A ne pas confondre avec la médiane, qui est la valeur qui scinde les effectifs en deux parties égales.

On va rechercher, comme pour le calcul d'une médiane, à quelle classe de salaire correspond la 159ème masse salariale (317+1)/2. Cette valeur est contenue dans la classe [8-10[. Par une interpolation linéaire on va trouver la médiale :

La formule s'écrit donc (équation de proportionnalité) $M = V + \frac{D}{E} A$ ©

ainsi dans l'interpolation la valeur cherchée (M) dépend de la valeur la plus faible de la classe (V=8), de la différence entre le rang de la valeur recherchée et les effectifs cumulés des classes précédentes (D=20), de l'effectif de la classe (E=90) et de l'amplitude de la classe (A=2). La médiale vaut 8.4.

3) On mesure l'écart entre la médiale et la médiane :

Cet écart vaut 8.4-7.5 = 0.9. Cet écart traduit la concentration. Si cet écart est grand par rapport à l'étendue du caractère, la concentration est forte, s'il est nul, la concentration est nulle, on est dans une égalité parfaite, dans cette hypothèse tous les salariés toucheraient le même salaire, la richesse est bien répartie.

4) Calculons IC. Dans cet exemple l'étendue vaut (12-2)=10, donc la concentration n'est pas très élevée (0.9/10*100) = 9%. Lorsque la concentration est forte, peu de salariés perçoivent une part importante de la masse salariale.

E. Exercice

Soit deux candidats X et Y qui se présentent à un concours comportant 5 disciplines (de A à E). Leurs résultats sont portés dans le tableau ci-dessous.

	A	B	C	D	E
X	5	14	7	12	12
Y	15	12	16	4	8

Quelle est l'homogénéité de leurs résultats ?

Calculer la moyenne de X : $\bar{x}$ = 50 / 5 = 10

Etablir le tableau suivant pour calculer l'écart-type :

						Somme
Note X	5	14	7	12	12	50
$x-\bar{x}$	-5	4	-3	2	2	0
$(x-\bar{x})^2$	25	16	9	4	4	58

On remarque que la somme des $x-\bar{x}$ vaut 0

$$\sigma = \sqrt{\frac{1}{5} \times 58} = 3,40$$

Calculer le coefficient de variation C : C = 3,4 / 10 = 0,34

Faire pareil pour le candidat Y :

Moyenne de Y : $\bar{y}$ = 55 / 5 = 11

$$\sigma = \sqrt{\frac{1}{5} \times 100} = 4,47$$

C = 4,47 / 11 = 0,40

Conclusion : le candidat Y a une moyenne supérieure à celle de X. Cependant ses résultats sont plus dispersés, aussi bien de façon absolue (écart-type), que de façon relative (coefficient de variation).

IV. CONSTRUCTION DES CLASSES : DISCRETISATION

Ce problème ne se pose que pour les variables quantitatives. Pour les variables qualitatives les classes correspondent à la nomenclature.
Discrétiser une variable quantitative se dit en langage courant "réaliser un découpage en classes". En statistiques, discrétiser c'est à la fois réaliser ce découpage, nommer et justifier les classes. Pour réaliser une discrétisation, il faut choisir le nombre de classes et les bornes de classe.
Qu'est-ce qu'une bonne discrétisation ? Intuitivement, un bon découpage correspond à des classes homogènes et séparées : les objets géographiques d'une même classe doivent se ressembler plus entre eux qu'ils ne ressemblent aux objets des autres classes.

A. Choix du nombre de classes

Le nombre de classes optimum à réaliser dans une partition est toujours fonction du nombre d'individus observés.
Il existe quelques formules "toute faites" pour déterminer à l'aveugle le nombre n de classes à partir du nombre N de données : ©

Brooks-Carruthers : $5*\log_{10}(N)$
Huntsberger : $1 + 3{,}332*\log_{10}(N)$
Sturges : $\log_2(N+1)$

Deux autres formules, censées être plus précises, mettent en jeu le minimum **a** des données et le maximum **b** et utilisent aussi d'autres paramètres de la dispersion : σ, l'écart-type et **eiq** l'écart interquartiles : ©

Scott : (b-a)/(3.5*sig*N**(-1/3))
Freedman-Diaconis : (b-a)/(2*eiq*N**(-1/3))

B. Choix des bornes de classes

Il existe de nombreuses méthodes, dont la plupart ont des critères explicites de découpage et des formules mathématiques pour calculer les bornes. La plupart de ces méthodes supposent que le nombre de classes a été fixé. Nous ne présentons ici que quelques méthodes.

1) La méthode des quantiles ("des effectifs égaux")

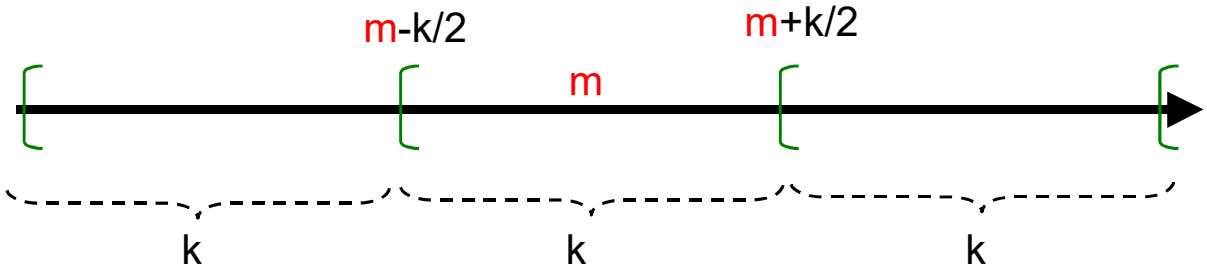
Le critère visé est l'équirépartition, c'est à dire le même nombre de données par classe. Dans la version stricte, à partir du nombre du nombre **N** de données et du nombre **n** classes, on en déduit le nombre **F** d'individus par classe. On trie les données par ordre croissant et on met dans la classe 1 les **F** premières données, dans la classe 2 les **F** suivantes etc.

Un trop grand nombre de valeurs égales perturbe la discrétisation, aussi, dans la version relâchée, on met éventuellement plus de **F** données par classe car on force les données égales à être dans une même. Voici ce que cela donne sur un exemple de 6 valeurs avec 2 classes :

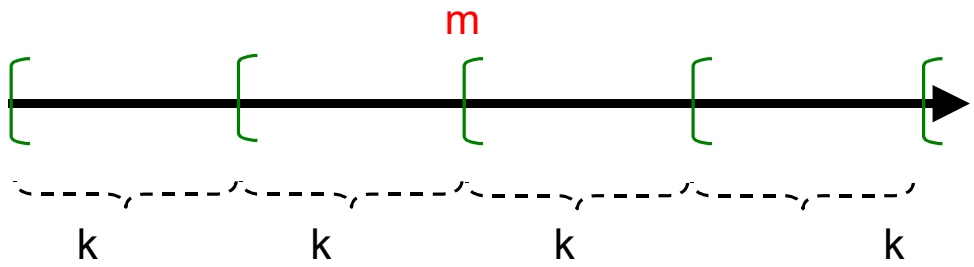
Données	:	10	11	12	12	13	14
Version stricte	:	1	1	1	2	2	2
Version relâchée	:	1	1	1	1	2	2

2) La méthode des amplitudes

On garantit ici que le critère d'égalité d'amplitude de classe est respecté, l'amplitude étant la différence entre la plus grande valeur et la plus petite valeur. A partir du minimum global **a** des données et du maximum global **b** des données on calcule les bornes de classe $\mathbf{h_i}$ à l'aide d'une simple progression arithmétique dont la raison est **k=(b-a)/(n-1)**. Une variante de cette méthode consiste à prendre comme largeur **k** la valeur de l'écart-type des données. Si **n** est impair, la classe du milieu a pour bornes **m-k/2,m+k/2** où **m** est la moyenne des données.



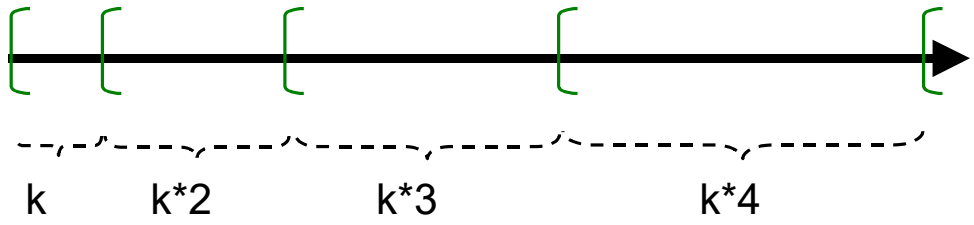
Si **n** est pair, **m** est la borne supérieure de la classe numéro **n/2**.



Cette méthode ne convient pas si la distribution des données est trop dissymétrique : les classes pourraient être très inégales et certaines vides !

3) Progression arithmétique ou géométrique

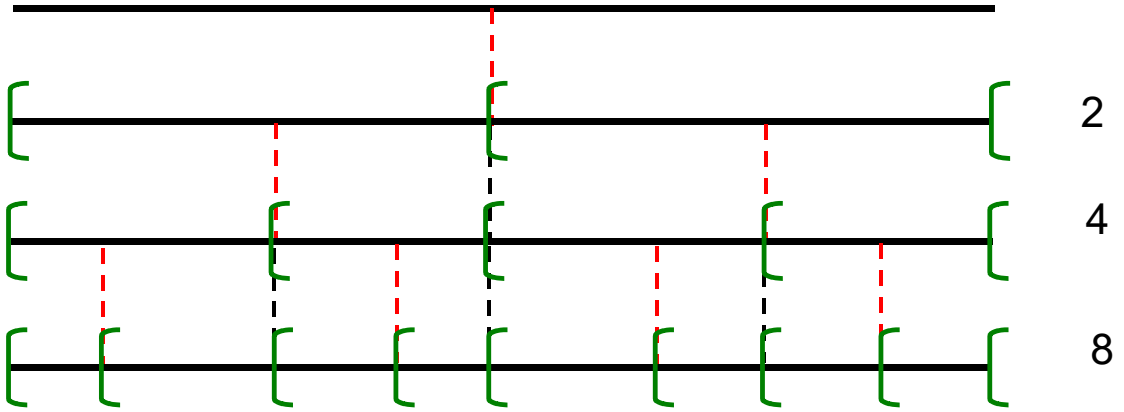
Les bornes supérieures des classes sont des multiples des classes précédentes ou calculées différemment.



L'intérêt est de mieux étaler la répartition dans les faibles valeurs, plus finement représentées, par contre les fortes valeurs se retrouvent regroupées dans la dernière classe. Les classes sont donc d'amplitude croissante. Ceci ne peut s'appliquer un un certain type de distributions

4) La méthode des moyennes emboitées

Le nombre de classes est ici une puissance de deux. On sépare l'intervalle de départ en deux en prenant comme valeur de séparation la moyenne globales des valeurs. On recommence ensuite en découpant chaque classe en deux en prenant comme valeur de séparation la moyenne des valeurs de la classe.



5) Standard

On utilise la moyenne et l'écart-type : toutes les classes ont la même amplitude, égale à l'écart-type. Pour un nombre impair de classe, la moyenne se situe à la médiane de la classe centrale, et pour un nombre pair de classes, la moyenne se situe à la limite des classes. Très bonne méthode.

6) La méthode des grandes différences relatives

On trie les valeurs par ordre croissant puis on calcule les différences relatives successives entre une valeur et sa précédente. On change de classe lorsque la différence relative est supérieure à un seuil arbitraire, classiquement 50 %. Le nombre de classes n'est donc pas fixé a priori.

x_1 x_2 x_3 x_4 … x_n

$dif = (x_i - x_{i-1}) / x_{i-1} * 100$

Si dif > 50 on change de classe.

7) Libre

Si la série a une distribution par paquets séparés par des coupures significatives

8) On veut deux classes :

On scinde l'effectif de part et d'autre de la médiane ou de la moyenne.