Statistiques

Présenté par:
A. Hadri
hadri.aissam@gmail.com

2023-2024

FPO

- Comparaison Mode-Moyenne-Médiane
- Coefficient d'aplatissement (Kurtosis)

4 Statistique à deux dimensions

- Tableau de contingence
- Distributions conditionnelles
  - Fréquences conditionnelles de X sachant Y
  - Fréquences conditionnelles de Y sachant X
- Représentation graphique
- Covariance
- Corrélation
- Coéfficient de corrélation linéaire
- Régression linéaire
- Méthode de Moindres carrés

Contents

Statistique descriptive

Introduction

La statistique est une étude approfondie d'un ensemble de données collectées. Cette étude est donnée sous forme d'analyse, traitement et interprétation des résultats ainsi que leur présentation afin de rendre les données plus accessible et compréhensibles pour un public non spécialiste. C'est à la fois une science de donnés, une méthode mathématique et un ensemble varié de techniques.

Statistique descriptive

Introduction

La démarche statistique

- Statistique descriptive : Son but est de synthétiser, résumer l'information contenue dans les données. Elle utilise les tableaux statistiques et les représentations graphiques.
- Statistique inférentielle : Son but est d'étendre les propriétés constatées sur l'échantillon à toute la population et valider des hypothèses a priori. Le calcul des probabilités joue un rôle important.

Dans la suite de ce cours, nous nous intéresserons au **statistique descriptive**.

Statistique descriptive

Introduction

On peut donc définir le mot statistique descriptive comme un ensemble de données et l'activité qui consiste à les recueillir, les traiter et les interpréter. Faire de la statistique c'est étudier un ensemble d'objets de même nature appelés individus ou unités statistique, sur lesquels on observe des caractéristiques appelées variables.

Statistique descriptive

Introduction

L'analyse des données est utilisée pour décrire les phénomènes étudiés, faire des prévisions et aider à prendre des décisions. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes.

Statistique descriptive

Introduction

Les phénomènes étudiés sont divers et de nature différentes, ce qui nécessite des études statistiques variées prenant en compte différents champs disciplinaires. Ceci explique pourquoi les cours de statistique interviennent dans plusieurs filières : Biologie, chimie, économie, ingénierie, mathématiques...

Statistique descriptive

Introduction

On peut donc énumérer les étapes d'une étude statistique comme suit :

1. La collecte des données.
2. Présenter et résumer ces données. En général sous forme de graphes.
3. Étude des paramètres et des indicateurs.
4. Tirer des conclusions sur le comportement de la population étudiée et prise de décision.
5. En se basant sur l'analyse des données, effectuer des prévisions.

Statistique descriptive

Introduction

L'analyse des données est donc définie comme l'ensemble des méthodes permettant une étude approfondie d'informations et de données de nature qualitative ou quantitative (qu'on définira dans la suite).

Statistique descriptive

Introduction

Dans l'analyse des données, on distingue :

- **L'analyse univariée**, qui porte sur l'étude d'une seule variable statistique (statistique à une seule variable).
- **L'analyse bivariée**, qui a pour objectif définir les relations entre deux variables d'une même population (statistique à deux variables).
- **L'analyse multivariée**, qui a pour but l'étude de plusieurs variables d'une population précise.

Statistique descriptive

Introduction

Pour élaborer des analyses statistiques, il est impératif de distinguer leurs types d'échelles (**échelles nominale, ordinale**,...), car les techniques et les méthodes utilisées dépendent continument de **la nature des variables auxquelles sont appliquées**. Ainsi par exemple pour la représentation graphique, on n'utilisera pas le même graphe pour présenter deux variables : nominale et ordinale ou deux variables quantitatives.
On commence par définir un peu de vocabulaire.

Statistique descriptive

Terminologie

Les statistiques font appel a diverses méthodes de classement des données tels que les tableaux, les histogrammes et les graphiques, permettant d'organiser un grand nombre de données. Ceci nécessite un vocabulaire assez particulier. On présentera ainsi c-joint quelques définitions de termes les plus utilisés dans l'étude statistique :

- Population : ensemble des unités sur lesquelles porte l'étude. elle constitue l'ensemble des individus (noté Ω).
- Échantillon : sous-ensemble d'unités de population.
- Individu ou Unité statistique : élément de cette population.
- Effectif total : nombre d'individus de la population. Noté par N

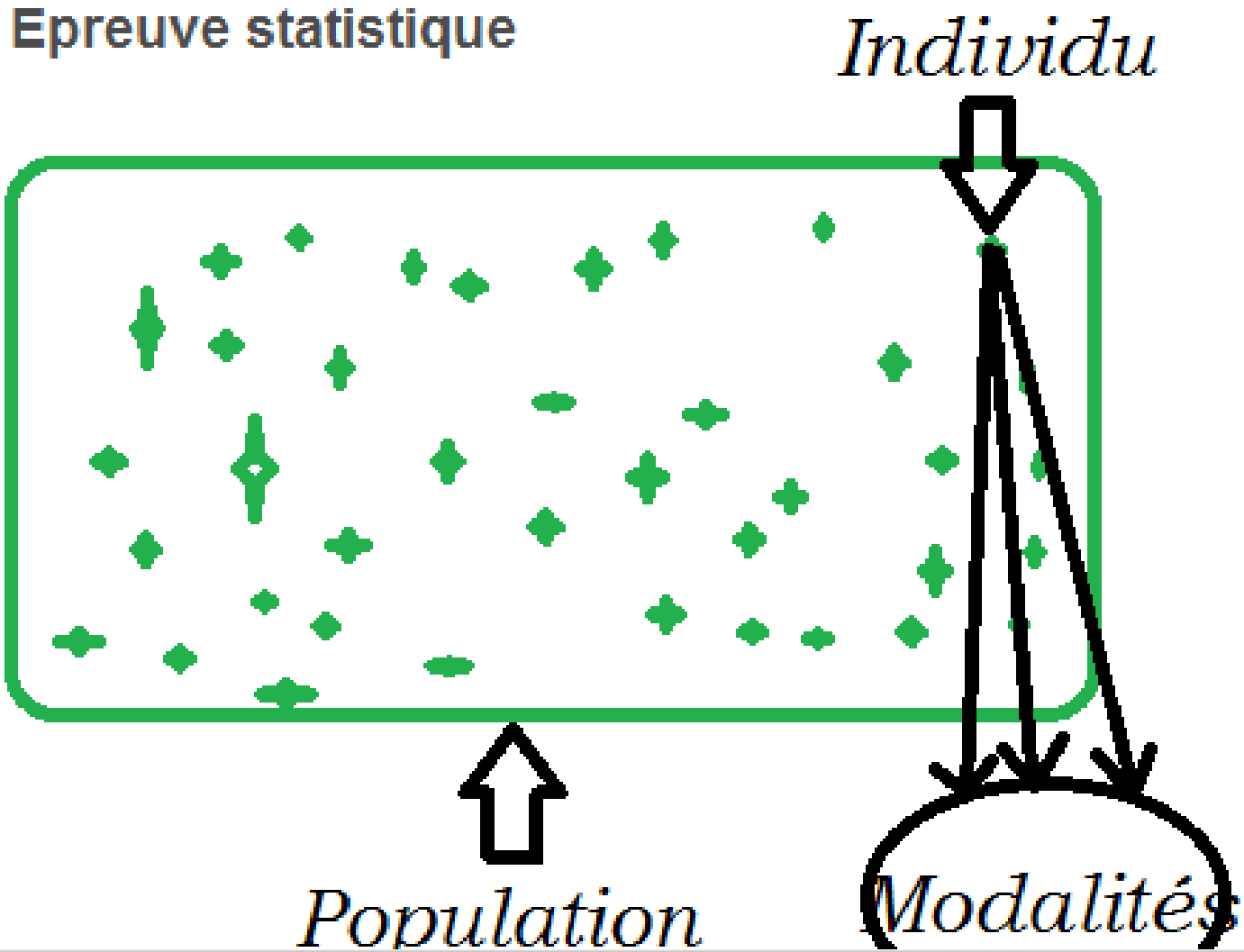
Statistique descriptive

Terminologie

- Variable statistique ou Caractère : caractéristique étudiée sur la population. C'est la caractéristique dont porte l'étude statistique. Noté par X.
- Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci. Notés par $(x_i)_{i=1\ldots N}$
- Série statistique : C'est la correspondance entre individu de la population et le caractère étudié. Noté par $\{(x_i)_{i=1\ldots N}\}$.
- Enquête : étude d'une ou de plusieurs variables sur une population donnée.

Statistique descriptive

Terminologie



Statistique descriptive

Exemple

On veut faire une étude sur le genre de 20 étudiants d'une classe CM1. On définit alors les différents termes de la série statistique comme suit :

- Population : Les 20 étudiants.
- Individu ou Unité statistique : Étudiant.
- Effectif total : $N = 20$
- Variable statistique ou Caractère : Le genre.
- Les modalités : Fille (F) ou Garçon (G).
- Série statistique : $F, F, G, G \ldots$.

Statistique descriptive

Statistique descriptive

Avant de définir les types du caractère, on énonce tout d'abord la définition de la statistique descriptive.

Definition (Statistique descriptive)

La statistique descriptive est un ensemble de méthodes basées sur la bonne collecte des données pour représenter graphiquement et résumer numériquement les données pour mieux décrire et analyser les comportements de la population.

Statistique descriptive

Types de caractère

Ainsi, il faut bien maitriser les types de caractère car chacun a une étude différente de l'autre.
On classe les variables selon **leur nature**. Nous distinguons en effet deux catégories de caractères, à savoir : **les caractères de type qualitatif et les caractères de type quantitatif**. On dit alors qu'une variable est

- **Qualitative** : si ses valeurs ne sont pas mesurables. Les modalités ne sont pas des valeurs numériques. De plus, on précise souvent si elle est
  - **Nominale**, lorsque les modalités ne peuvent pas être ordonnées.
  - **Ordinal**, lorsque les modalités sont ordonnées selon une certaine hiérarchie.

Statistique descriptive

Types de caractère

- **Quantitative** : si elle prend des valeurs numériques. De plus, on précise souvent si la variable quantitatif est
  - **Discrète**, lorsqu'elle prend des valeurs entières (discrètes).
  - **Continue**, lorsqu'elle prend des valeurs réelles (en général sous forme d'intervalles).

Statistique descriptive

Variables qualitatives

Les caractères qualitatifs sont ceux dont **les modalités ne peuvent pas être ordonnées**, c'est-à-dire que si l'on considère deux caractères pris au hasard, **on ne peut pas dire que l'un des caractères est inférieur ou égal à l'autre**. Donc pour définir ce type de caractère il faut voir les modalités de la variable. Voici quelques exemples du caractère qualitatif :
examlpe :
Les villes, la couleur, le genre, la nationalité, la religion, lieu de naissance, la fréquence d'une activité (jamais, parfois, souvent, très souvent).

Statistique descriptive

Variables qualitatives

On distingue alors comme on a précisé avant deux sous familles du caractère qualitatif :

- **Les variables qualitatives nominales** : il n'y a pas de hiérarchie entre les différentes modalités ; exemple : sexe, couleur des yeux, nationalité, villes, lieu du travail, les branches, la langue...
- **Les variables qualitatives ordinales** : les différentes modalités peuvent être ordonnées suivant une certaine logique ; exemple : la mention au baccalauréat, la qualité du service, l'état d'une voiture, salariés...

Statistique descriptive

Variables qualitatives

Parfois, on peut passer d'un type de caractère à un autre en donnant par exemple une valeur numérique à une modalité (nominale), où bien l'inverse.

- Certaines variables nominales (ordinales) peuvent être remplacés par un code numérique, qui n'a pas de valeur de quantité. Exemple : le code postal (40000—> Marrakech), la taille (1.90—> haute, 170—> moyenne, 1.50—> petite)le sexe (1—>garçon, 2—>fille).
- Si une variable qualitative ne contient que deux modalités, on l'appelle variable dichotomique.

Statistique descriptive

Variables qualitatives

Supposons qu'on ait demandé à un groupe composé de dix étudiants pris au hasard de l'université. On leur pose la question sur la matière préférée.

- Effectif total = 10.
- Population : groupe de 10 étudiants.
- Individu : étudiant du groupe
- Caractère : matière préférée
  - valeurs (modalités) : Chimie (C), Physique (P), Mathématiques (M), Informatique (I), Biologie (B).
  - Type du caractère : variable qualitative nominale.

Statistique descriptive

Variables qualitatives

Données individuelles :

$C; \square\, C\ \square\, C\ \square\, C\ \square\, P\ \square\, P\ \square\, P\ \square\, I\ \square\, B\ \square\, B.$

- L'effectif associé à C est 4 car quatre étudiants ont répondu Chimie.
- L'effectif associé à P est 3 car trois étudiants ont répondu Physique.
- L'effectif associé à M est 0 car aucun étudiants n'a répondu Mathématiques.
- L'effectif associé à I est 2 car deux étudiants ont répondu Informatique.
- L'effectif associé à B est 1 car un étudiant a répondu Biologie.

Statistique descriptive

Variables qualitatives

On regroupe ces résultats dans un tableau appelé le tableau statistique. On rappel que les tableaux statistiques consistent à résumer et présenter les données observées sous forme numérique.

Tableau de distribution des effectifs selon la matière choisie sous forme horizontale

Réponse	C	P	M	I	B
Effectif	4	3	0	1	2

Statistique descriptive

Variables qualitatives

ou bien

Tableau de distribution des effectifs selon la matière choisie sous forme verticale

Réponse	Effectif
C	4
P	3
M	0
I	1
B	2

Souvent en statistique descriptive, il est préférable de considérer des pourcentages ou proportions pour avoir une bonne indication de la représentation de chaque modalité plutôt que de considérer le nombre d'individus discret. Cette proportion est ce qu'on appelle la **Fréquence**.

Statistique descriptive

Variables qualitatives

Definition (Fréquence)

Pour chaque valeur (ou modalité) x_i, on définit la **Fréquence** comme étant le rapport :

$$f_i = \frac{n_i}{N},$$

c'est la proportion des individus de la population ayant cette modalité.

Remarque :

- Une fréquence est soit exprimée en pourcentage (par exemple 30%) soit par un nombre compris compris entre 0 et 1 (par exemple 0.25).
- La fréquence est toujours comprise entre 0 et 1.
- On a toujours $\Sigma_i f_i = 1$.

Statistique descriptive

Variables qualitatives

Par exemple si on veut déterminer la fréquence des étudiants qui ont répondu par Chimie, elle est donnée par :

$$f_1 = \frac{4}{10} = 0.4 \, ou \, 40\%$$

On calcule de la même façon les proportions de P, M, I et B. On regroupe alors tous les résultats dans un tableau, appelé **tableau des proportions ou fréquences**

Réponse	C	P	M	I	B
Fréquence	0.4	0.3	0	0.1	0.2

Statistique descriptive

Variables qualitatives

Remarque Le traitement statistique d'une variable qualitative est très limité. En effet, il existe peu de grandeurs ou indicateurs dont on peut s'inspirer pour étudier une distribution statistique. Le moyen le plus crédible est de représenter graphiquement les données. Par contre, aucune étude paramétrique sera envisagé après.

Statistique descriptive

Variables qualitatives

On représente graphiquement une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative). Les représentations recommandées et les plus fréquentes sont les diagrammes ou graphes. Dans la suite nous étudierons la représentation graphique d'une série statistique à caractère qualitatif.

Statistique descriptive

Représentation graphique

Dans la littérature, de nombreuses représentations plus ou moins informatives peuvent être utilisées. Parmi ces représentations, on trouve les deux diagrammes assez classiques permettent de représenter une variable qualitative : **le diagramme en bandes (dit tuyaux d'orgue)** et **le diagramme à secteurs angulaires (dit camembert)**.

Statistique descriptive

Les tuyaux d'orgue

Pour présenter les données sous forme de tuyaux d'orgue, il faut suivre les deux instructions suivantes :

- les modalités de la variable sont placées sur une droite horizontale (attention : ne pas orienter cette droite car les modalités ne sont pas mesurables et il n'y a donc pas de relation d'ordre entre elles).
- les effectifs (ou les fréquences) sont placées sur un axe vertical. La hauteur du tuyau est proportionnelle à l'effectif.

Statistique descriptive

Les tuyaux d'orgue

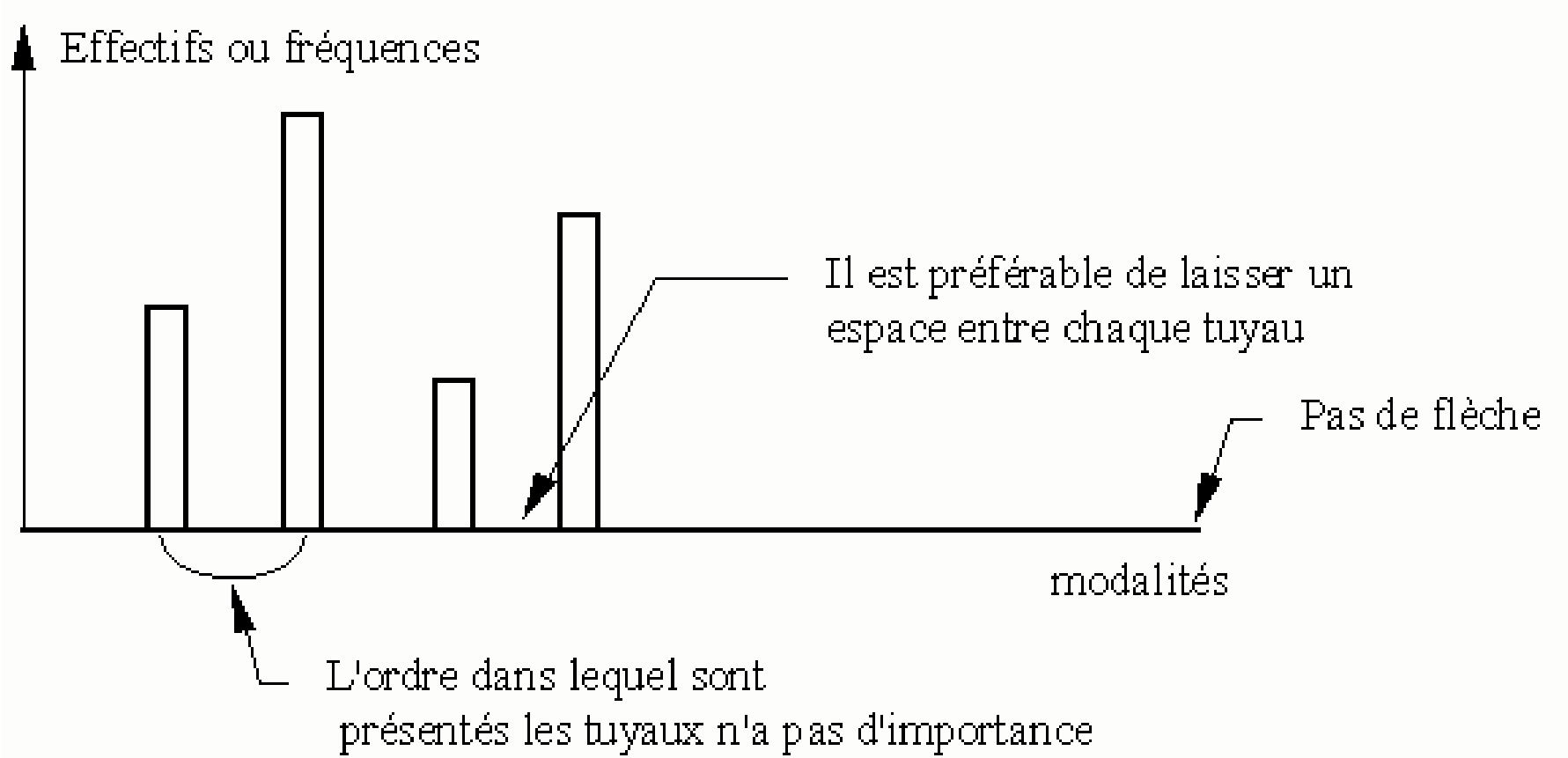


Figure – Le diagramme tuyaux d'orgue.

Statistique descriptive

Exemple

On reprend l'exemple de tout à l'heure des 10 élèves.

Tableau de distribution des effectifs

Réponse	C	P	M	I	B
Effectif	4	3	0	1	2

Statistique descriptive

Exemple

La représentation graphique associée à cette distribution est donnée par un diagramme tuyaux d'orgue (voir Figure 4).

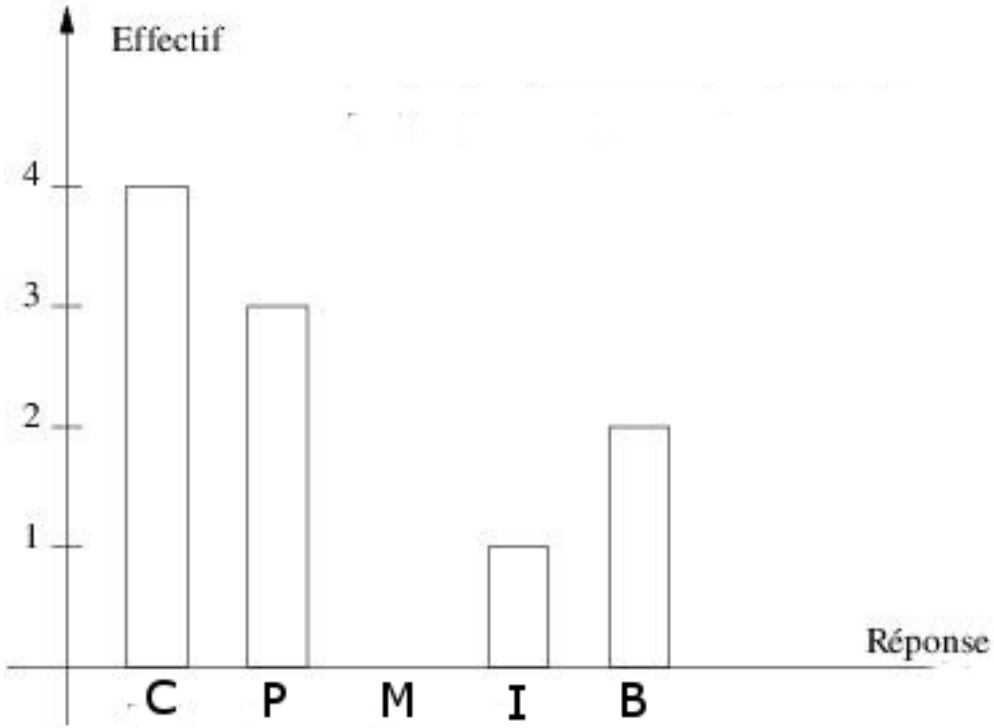


Figure – La matière préferée des 10 étudiants par effectif.

Statistique descriptive

Exemple

Tableau de distribution des fréquences.

Réponse	C	P	M	I	B
Fréquence	0.4	0.3	0	0.1	0.2

Statistique descriptive

Exemple

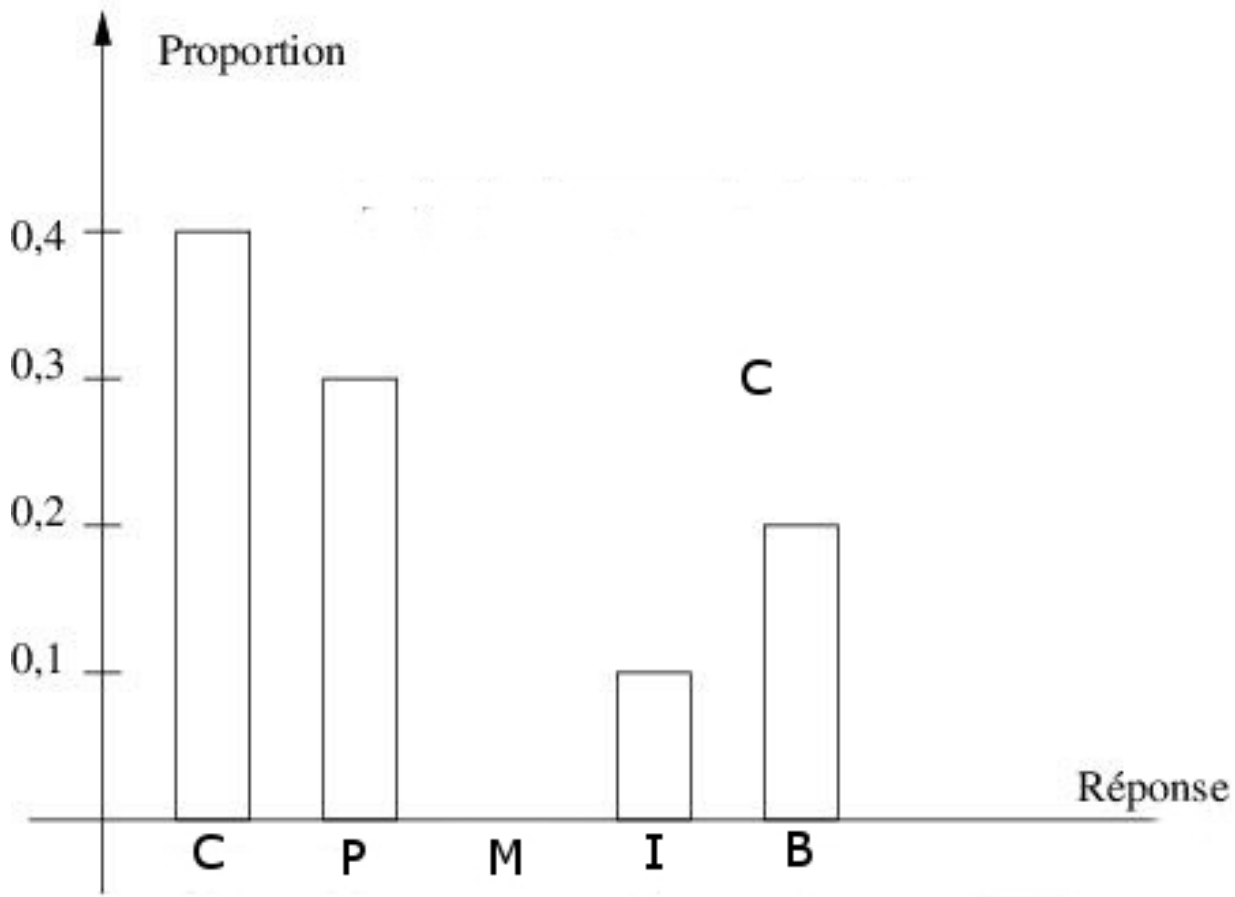


Figure – La matière préferée des 10 étudiants par fréquence.

Statistique descriptive

Remarque

Les tuyaux ont une certaine épaisseur pour qu'il n'y ait pas de confusion avec les diagrammes en bâtons réservés à la variable quantitative discrète.

Statistique descriptive

Définition

Diagramme par secteur (diagramme circulaire) : Les diagrammes circulaires, ou semi-circulaires, consistent à répartir un disque ou un demi-disque, en secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence de la modalité.
Pour présenter les données statistiques sous forme de diagrammes à secteurs (ou camemberts), il faut suivre les deux instructions suivantes :
- L'effectif total est représenté par un disque.
- Chaque modalité est représentée par un secteur circulaire dont la surface (pratiquement : l'angle au centre) est proportionnelle à l'effectif correspondant.

Diagramme par secteur

Exemple

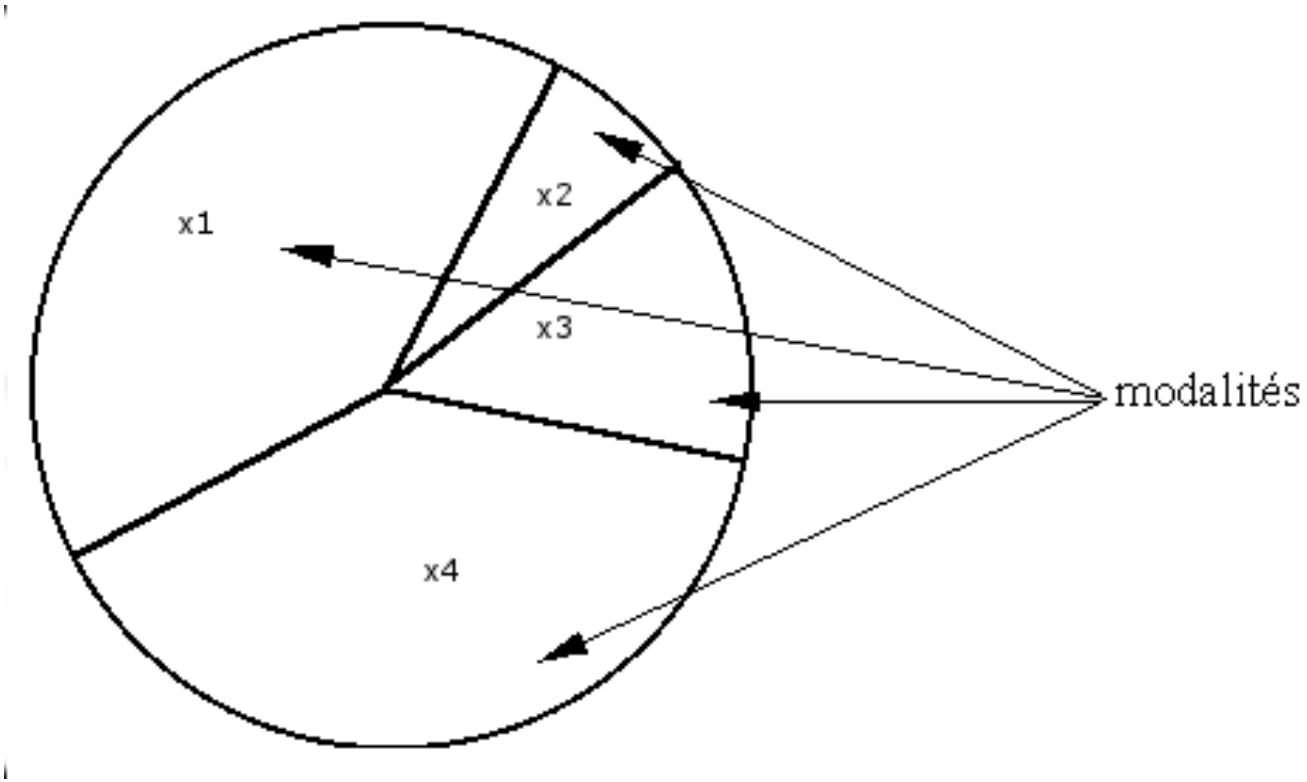


Figure – La représentation d'un diagramme circulaire.

Diagramme par secteur

La règle

En effet, le degré d'un secteur est déterminé à l'aide de la règle de trois, de la manière suivante :

$$N \longrightarrow 360^\circ$$

$$n_i \longrightarrow d_i \text{ (le degrés de la modalité } x_i$$

On trouve alors l'angle représentatif de chaque modalité par la relation suivante :

$$d_i = \frac{n_i \times 360}{N}.$$

Variables quantitatives discrètes

Définition

Le caractère statistique peut prendre un **nombre fini dénombrable de valeurs** (note, nombre d'enfants, nombre de couple, nombre de pièces, ...). Dans ce cas, le caractère statistique étudié est alors appelé un **caractère discret**. On traite ce type de variable comme une variable qualitative si elle prend un nombre assez petit de valeurs et comme une variable quantitative continue si elle prend un nombre considérable de valeurs.

Variables quantitatives discrètes

Exemple

Une société américaine a recensé le nombre d'enfants des familles d'un village au Maroc. Les données obtenues ont été collectées dans le tableau de distribution des effectifs suivant :

Tableau de distribution des effectifs

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

Variables quantitatives discrètes

Exemple

- Population : Les familles du village.
- Individu : une famille.
- Effectif total : 500+250+175+75=1000 familles.
- Variable : nombre d'enfants d'une famille.

On peut aussi considérer le tableau de distribution des fréquences (proportions) :
Par exemple $f_3 = \frac{175}{1000} = 0.175$ ou $f_3 = 17.5\%$

Variables quantitatives discrètes

Exemple

Tableau de distribution des fréquences

Nombre d'enfants	1	2	3	4
Fréquence	0.5	0.25	0.175	0.075

Variables quantitatives discrètes

Représentation graphique

Avant toute tentative de représentation, il faut tout d'abord distinguer entre variable discrète et variable continue (regroupements en classes).

Lorsque les modalités d'une variable discrète sont trop nombreuses, il est préférable de regrouper des modalités pour obtenir une variable classée afin que les graphiques synthétisent l'information et restent lisibles.
On passe alors du cas discret au cas continu (les classes ou intervalles). Une variable continue est systématiquement synthétisée dans un tableau sous forme de variables classées.

Variables quantitatives discrètes

Représentation graphique

Deux types de graphiques sont intéressants à représenter :

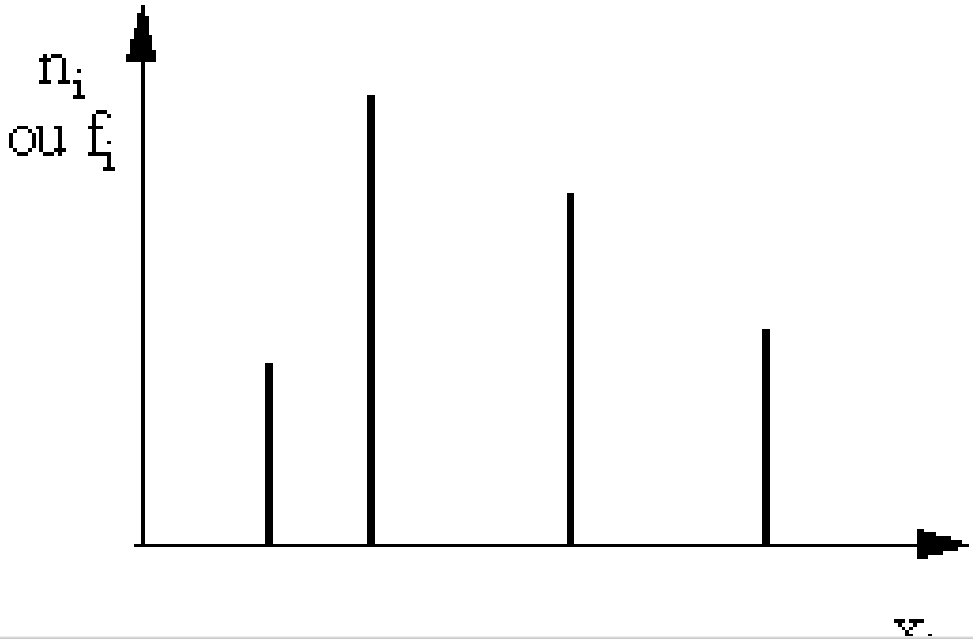
a) Les diagrammes différentiels qui mettent en évidence les différences d'effectifs (ou de fréquences) entre les différentes modalités ou classes.

b) Les diagrammes cumulatifs qui permettent de répondre aux questions du style "combien d'individus ont pris une valeur inférieure (ou supérieure) ?".

Variables quantitatives discrètes

Représentation graphique

Pour représenter une variable quantitative discrète on peut utiliser le diagramme différentiel appelé **diagramme en bâtons**. Les valeurs discrètes x_i prises par les variables sont placées sur l'axe des abscisses, et les effectifs (ou les fréquences) sur l'axe des ordonnées. La hauteur du bâton est proportionnelle à l'effectif (voir Figure 6).



Représentation graphique

Exemple

On retourne à l'exemple de tout à l'heure, où un recensement des nombre d'enfants dans des familles d'un village est étudié. Voici le tableau statistique correspondant :

Tableau de distribution des effectifs

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

Représentation graphique

Exemple

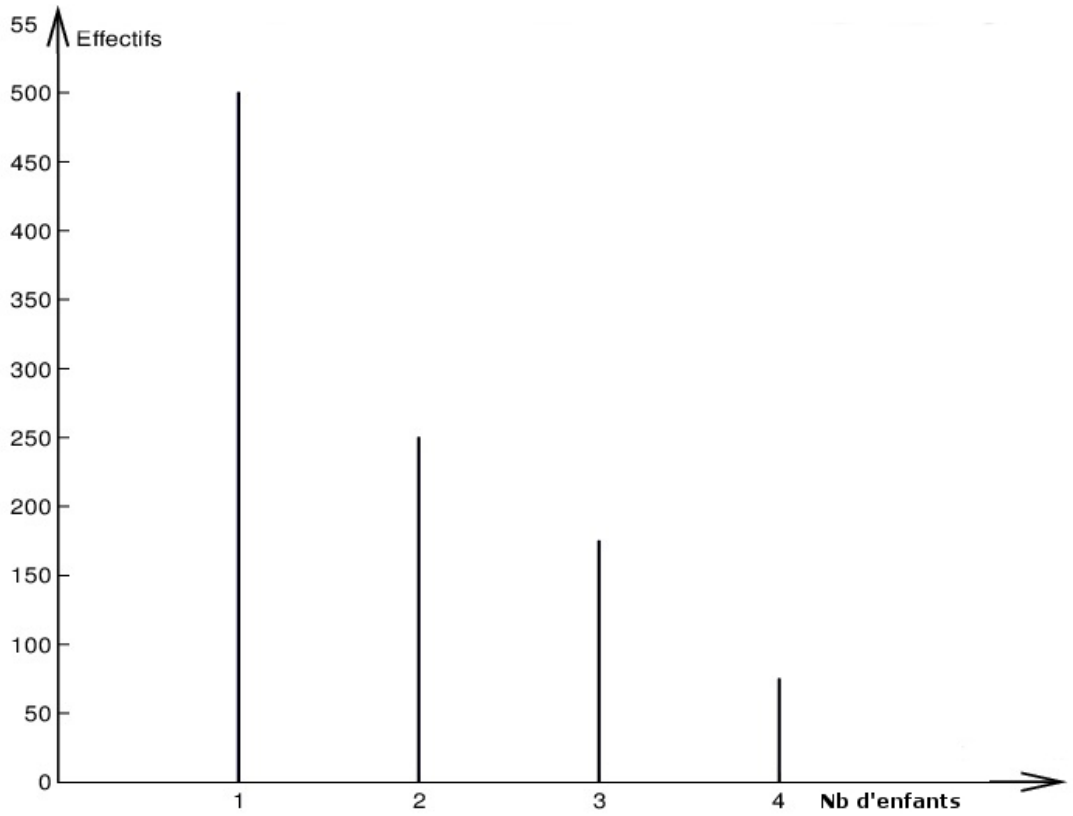


Figure – Le nombre d'enfants en fonction des familles du village en effectif.

Représentation graphique

Exemple

De la même façon, on présente le tableau des fréquence ainsi que le diagramme en bâton associé.

Tableau de distribution des fréquences

Nombre d'enfants	1	2	3	4
Fréquence	0.5	0.25	0.175	0.075

Représentation graphique

Exemple

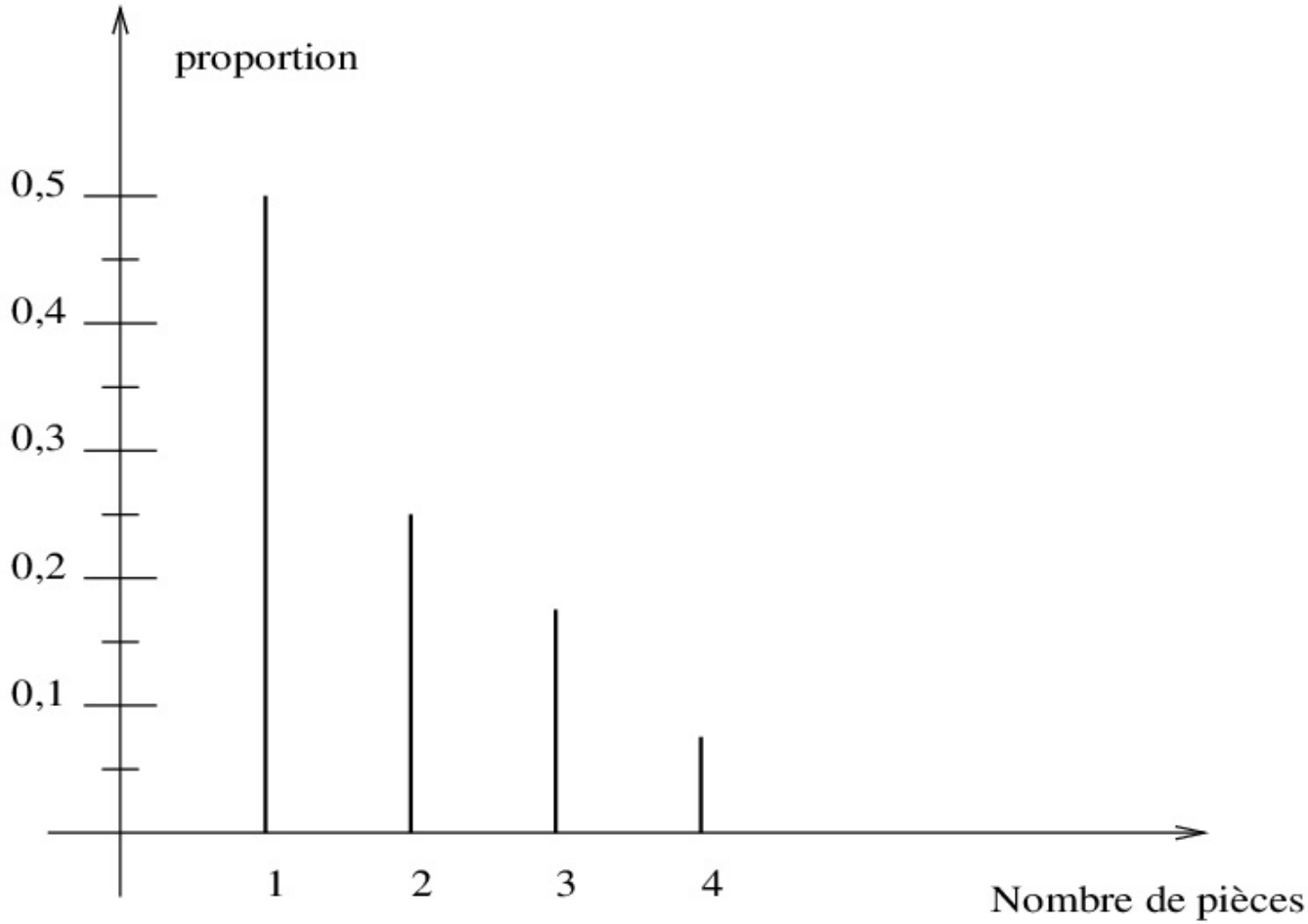


Figure – Le nombre d'enfants en fonction des familles du village en fréquence.

Variables quantitatives continues

Définition

Nous rappelons qu'une variable statistique quantitative continue concerne une grandeur mesurable sur des intervalles. Ses valeurs sont des nombres exprimant une quantité et sur lesquelles les opérations arithmétiques (addition, multiplication, etc,...) ont un sens. On peut donc parler des valeurs avec virgule.

En général, une variable continue est utilisée lorsque le nombre des modalités est assez grand, ce qui justifie le fait d'utiliser les classes.
Comment alors on détermine les classes ? Et comment on arrange les différentes valeurs dans ces classes.

Variables quantitatives continues

Définition

Voici les étapes qu'il faut suivre :

- On détermine dans un premier temps la plus petite et la plus grande valeur prise par la variable. Puis on choisi un nombre d'intervalles appelés classes de la forme $[x_i;\ x_{i+1}[$ (un intervalle semi-ouvert) couvrant l'ensemble des valeurs prises par la variable.
- Ensuite, pour chaque intervalle $x_i;\ x_{i+1}[$ on compte le nombres d'individus pour lesquels la variable prend une valeur supérieure ou égal à a et strictement inférieure à x_{i+1}. On appelle ce nombre, noté $n_{[x_i;\ x_{i+1}[}$, l'effectif de la classe $[x_i;\ x_{i+1}[$.
- Finalement, on regroupe dans un tableau à deux lignes ou à deux colonnes les différentes classes et leurs effectifs correspondants.

Variables quantitatives continues

Exemple

Un sondage est réalisé pour connaitre la durée en minutes du trajet domicile-FSTBM de dix étudiants de la FSTBM pris au hasard. Nous exprimons tout d'abord l'étude statistique :

- Population : 10 étudiants de FSTBM
- Variable statistique : durée (exprimée en minutes) du trajet domicile-FSTBM.
- Type de variable : quantitative continue. Données individuelles : $5 \square 6 \square 7 \square 10 \square 12 \square 13 \square 20 \square 25 \square 29 \square 39$
- La plus petite valeur = 5
- La plus grande valeur = 39
- On propose alors le choix des classes $[5;\ 15[$, $[15;\ 30[$ et $[30;\ 40[$

Variables quantitatives continues

Exemple

- L'effectif de la classe $[5;\ 15[$ est $n_{[5;\ 15[} = 6$ car six étudiants mettent moins de 15 minutes pour rejoindre la FSTBM.
- L'effectif de la deuxième classe $[15;\ 30[$ est $n_{[15;\ 30[} = 3$ car trois étudiants mettent plus de 15 minutes dans le trajet mais strictement inférieur à 30 minutes pour rejoindre la FSTBM.
- L'effectif de la dernière classe $[30;\ 40[$ est $n_{[30;\ 40[} = 1$ car un étudiant a besoin de plus de 30 minutes pour rejoindre la FSTBM.

Variables quantitatives continues

Exemple

On présente alors le tableau statistique en classe représentant la durée du trajet pour arriver à la FSTBM pour les 10 étudiants comme suit :

Tableau de distribution des effectifs

Durée	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Effectif	6	3	1

Variables quantitatives continues

Remarque

- La somme des effectifs des différentes classes doit être égal à l'effectif total.
- Le tableau de distribution des effectifs contient moins d'informations que les données individuelles.
  En effet, connaître l'effectif d'une classe ne donne pas d'information sur par exemple la répartition des données individuelles à l'intérieur de la classe.

Variables quantitatives continues

Remarque

Attention : Le choix des classes est très important ! En effet plus l'amplitude des classe est grande moins on a des informations concrètes sur la répartition des données.

Le choix du nombre de classes peut se faire selon la règle de Sturges :

$$Le\ nombre\ du\ classes\ = 1 + \log_2(N)$$

où N désigne l'effectif total.

Mais ce choix n'est pas toujours efficace surtout si les données ne sont pas bien réparties sur les classes.

Variables quantitatives continues

Remarque

On présente ci-dessous un peu de vocabulaire :

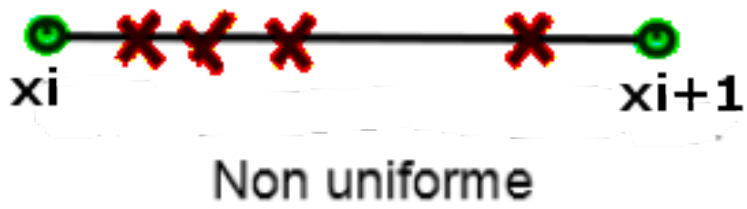
Vocabulaire :

- On appelle amplitude de la classe $[x_i;\ x_{i+1}[$, la quantité $a_i = x_{i+1} - x_i$.
- On appelle centre de la classe $[x_i;\ x_{i+1}[$, la quantité $c_i = \frac{x_i + x_{i+1}}{2}$.
- x_i est la borne inférieur de la classe $[x_i;\ x_{i+1}[$.
- x_{i+1} est la borne supérieur de la classe $[x_i;\ x_{i+1}[$.

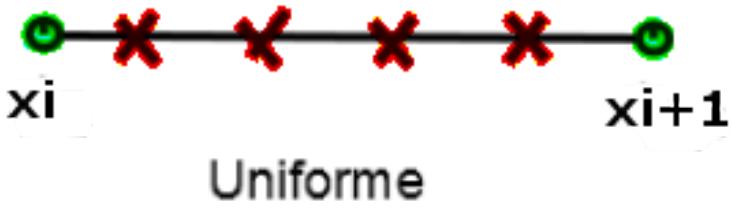
Variables quantitatives continues

Remarque

On supposera dans toutes les études qui suivent que la distribution à l'intérieur des classes est uniforme (voir Figure 9). Cette hypothèse permet de justifier le fait qu'on choisisse le centre des classes comme une moyenne des extrémités.



Non uniforme



Uniforme

Variables quantitatives continues

Exemple

- L'amplitude de la classe $[15;\ 30[$ est égale à $30 - 15 = 15$
- Le centre de la classe $[15;\ 30[$ est donné par $\frac{30+15}{2} = 22,5$

Variables quantitatives continues

Représentation graphique

Nous pouvons représenter le tableau statistique d'une variable statistique par un **Histogramme**. Nous posons les classes sur l'axe des abscisses et, au-dessus de chacune d'elles, nous traçons un rectangle dont l'aire est proportionnelle à l'effectif n_i (ou la fréquence f_i) associée. En d'autre termes, c'est un ensemble de rectangles contigus, chaque rectangle associé à chaque classe ayant une surface proportionnelle à l'effectif (fréquence) de cette classe.

Variables quantitatives continues

Représentation graphique

Attention : Avant toute construction d'histogramme, il faut faire attention et regarder si les classes sont d'amplitudes égales ou inégales. Si les classes sont de même amplitude, on représente directement les classes avec hauteur proportionnelle à l'effectif (ou fréquence). **Sinon il faut procéder par une correction de l'effectif ou bien une détermination préalable des densités de fréquence des différentes classes.**

Variables quantitatives continues

Représentation graphique

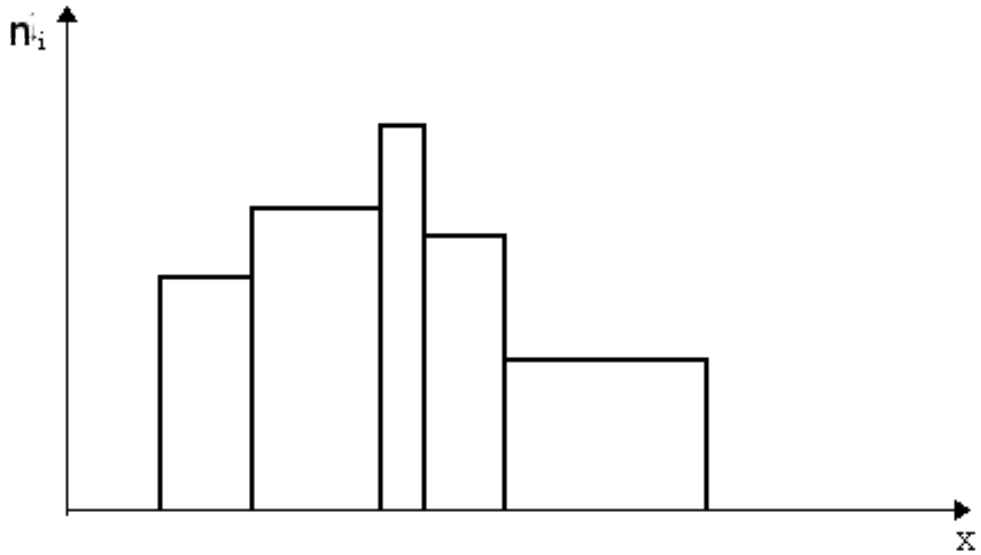


Figure – Une représentation d'un Histogramme.

Variables quantitatives continues

Représentation graphique

Dans un premier temps, nous utilisant les densités de fréquence des différentes classes pour tracer le Histogramme lorsque les classes ne sont pas de même amplitude.

Variables quantitatives continues

Densité de fréquence

Étant donnée une classe $[x_i;\ x_{i+1}[$ de fréquence $f_{[x_i;\,x_{i+1}[} = \frac{n_{[x_i;\,x_{i+1}[}}{N}$. La densité de fréquence de la classe (notée d_i) $[x_i;\ x_{i+1}[$ est alors donnée par $d_i = \frac{f_{[x_i;\,x_{i+1}[}}{x_{i+1}-x_i}$.

Pour mieux comprendre comment il faut tracer l'histogramme pour les classes qui n'ont pas la même amplitude, on considère l'exemple suivant :

Variables quantitatives continues

Densité de fréquence

Durée du trajet domicile-école de 10 étudiants.

Tableau de distribution des proportions

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Fréquence	0.6	0.3	0.1
Amplitude	10	15	10

Variables quantitatives continues

Densité de fréquence

On commence par calculer les densités pour chaque classe :

La densité de fréquence de la classe $[5;\ 15[$ est donnée par $d_1 = \frac{0,6}{10} = 0.06$

La densité de fréquence de la classe $[15;\ 30[$ est donnée par $d_2 = \frac{0,3}{15} = 0.02$

La densité de fréquence de la classe $[30;\ 40[$ est donnée par $d_3 = \frac{0,1}{10} = 0.01$

On résume alors les densités calculées dans le tableau suivant :

Variables quantitatives continues

Densité de fréquence

Classe (x_i)	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Densité de fréquence (d_i)	0.06	0.02	0.01

Variables quantitatives continues

Densité de fréquence

Classe (x_i)	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Densité de fréquence (d_i)	0.06	0.02	0.01

Variables quantitatives continues

Densité de fréquence

On passe maintenant à la représentation graphique.

Histogramme

Pour représenter une variable quantitative continue, on utilise le histogramme.

On positionne sur l'axe des abscisses les différentes classes $[x_i;\ x_{i+1}[$.

Ensuite, on trace des rectangle dont la longueur est proportionnelle à la densité de fréquence correspondante à ces classes. Ci-dessous le tableau et le histogramme associé à l'exemple en haut.

Variables quantitatives continues

Exemple

Durée du trajet domicile-école

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Densité de fréquence	0.06	0.02	0.01

Variables quantitatives continues

Exemple

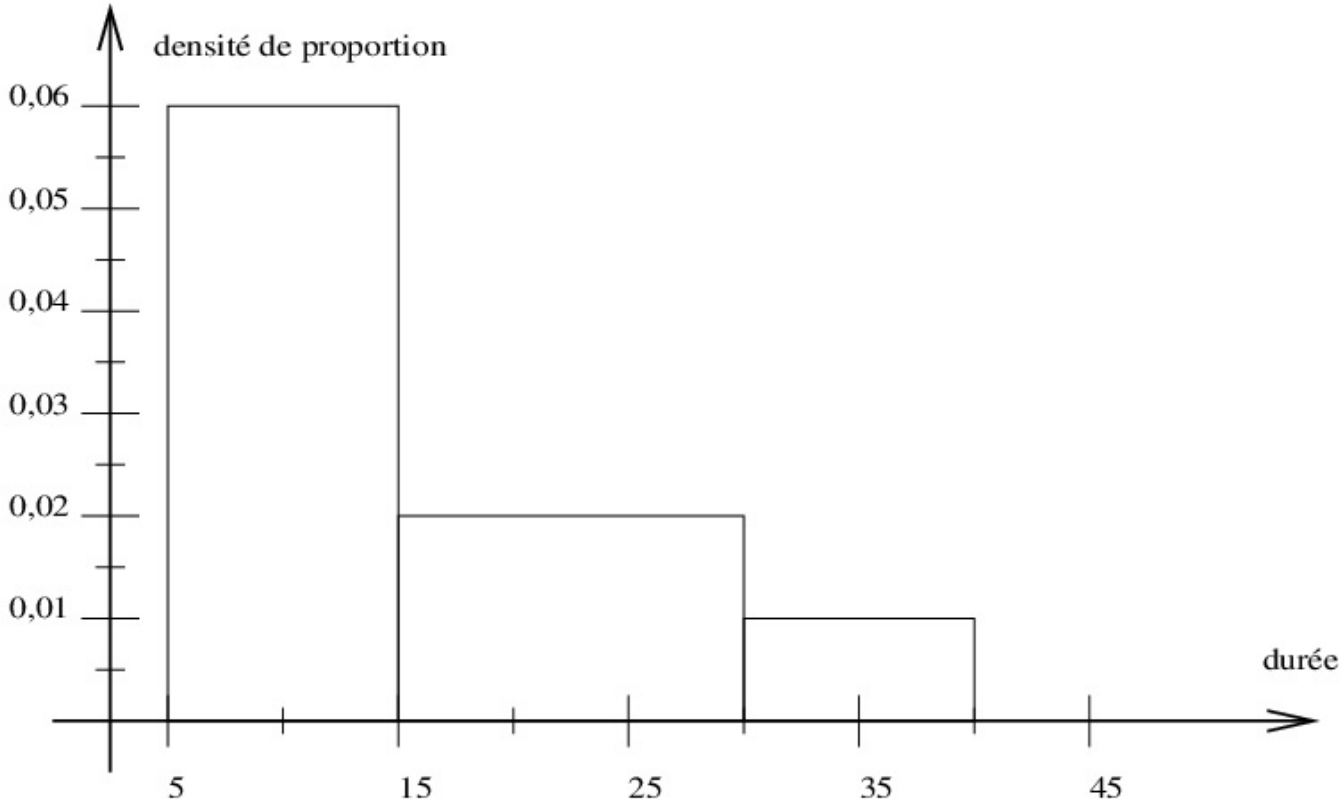


Figure – Le histogramme correspondant aux densités de fréquence.

Variables quantitatives continues

Exemple

Ainsi, la surface de chaque rectangle est exactement la fréquence de la classe correspondante.
On peut même le démontrer facilement, soit S la surface du rectangle dans l'histogramme, on a alors : $S =$ hauteur $\times$ largeur = $\frac{f_{[x_i;\, x_{i+1}[}}{x_{i+1}-x_i} \times (x_{i+1} - x_i) = f_{[x_i;\, x_{i+1}[}$.

La surface totale de l'histogramme est égale à 1 puisqu'elle est la somme des surfaces des rectangles dont leur surface et les fréquences correspondantes.

Variables quantitatives continues

Exemple

On revient à l'exemple précédent. La fréquence de la classe $[5;\ 15[$ est exactement la surface du rectangle correspond dont la surface est égale

$$S_{[5;\ 15[} = 0,06 \times (15 - 5) = 0,6.$$

Parfois on ne dispose pas du tableau statistique mais à la place on a un histogramme. Dans ce cas il faut savoir estimer la fréquence à partir de l'histogramme.

Variables quantitatives continues

Estimation d'une fréquence à partir de l'histogramme

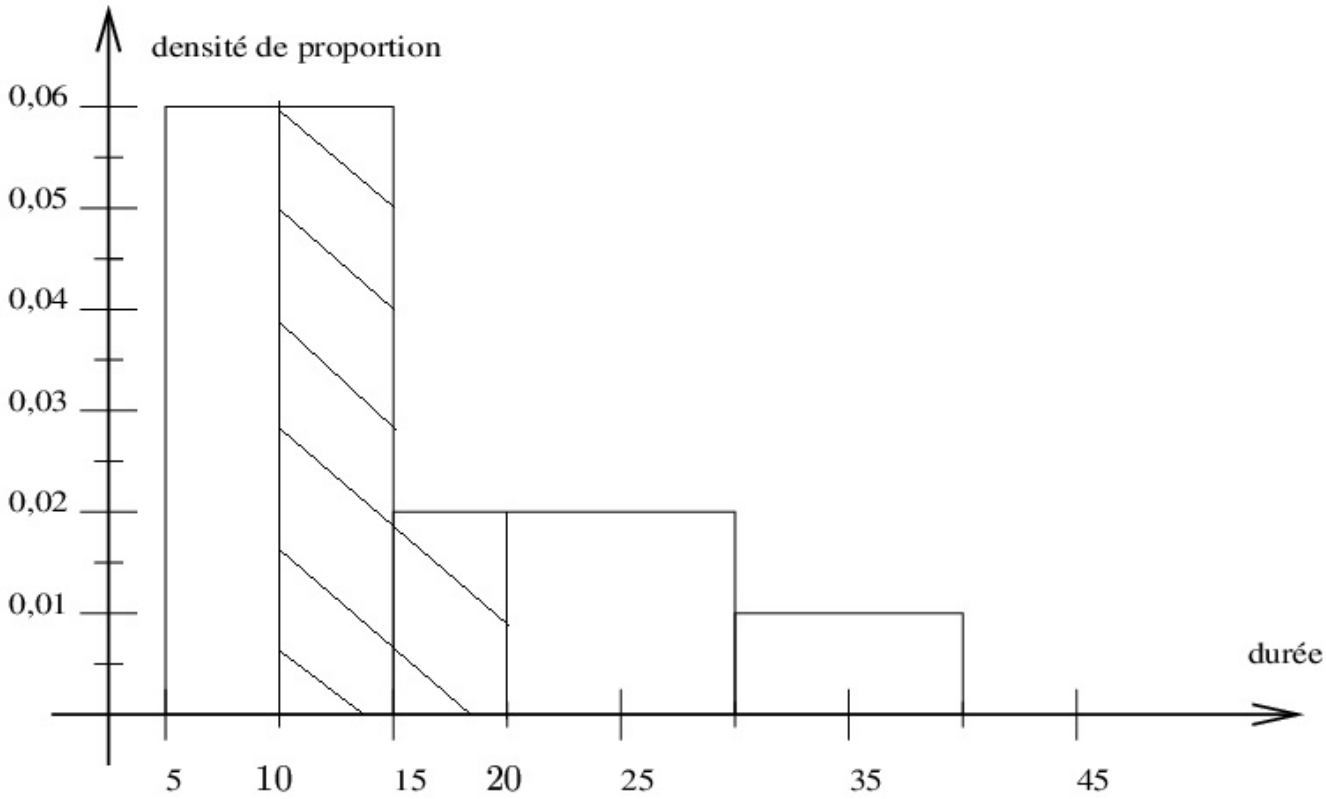
On reprend toujours l'exemple de tout à l'heure. L'objectif est de déterminer la fréquence des données comprises entre 10 et 20. Si on revient à l'histogramme correspondant à cet exemple, la fréquence est exactement la surface de l'air hachurée. Le calcul de cette aire (notée $S_{[10;\,20[}$ est donné comme suit :

$$S_{[10;\,20[} = [0.06 \times (15 - 10)] + [0.02 \times (20 - 15)] = 0.3 + 0.1 = 0.4.$$

Variables quantitatives continues

Estimation d'une fréquence à partir de l'histogramme

Donc, 40% des élèves on besoin de 10 à 20 minutes pour y arriver à l'école.



Variables quantitatives continues

Répartitions des effectifs et des fréquences

Cas discret : Soit X une variable quantitative discrète définie dans une population Ω, prenant les valeurs $x_1; \ldots; x_p$ classées par ordre croissant avec les effectifs partiels $n_1; \ldots; n_p$ respectivement.

Variables quantitatives continues

Effectifs cumulés

Définition :

On appelle effectif cumulé croissant de la modalité x_k la somme des effectifs partiels des valeurs $x_1; \ldots; x_k$. c'est à dire

$$N_k = n_1 + \ldots + n_k = \sum_{i=1}^{k} n_i$$

Variables quantitatives continues

Effectifs cumulés

Ainsi, l'effectif cumulé croissant associé à la modalité x_i du caractère X est le nombre d'individus de la population Ω dont les modalités sont inférieures ou égales à x_i.

Variables quantitatives continues

Effectifs cumulés

Definition

On appelle effectif cumulé décroissant de la valeur x_k le nombre

$$\tilde{N}_k = N - (n_1 + \ldots + n_{k-1}) = N - \sum_{i=1}^{k-1} n_i.$$

On considère l'exemple suivant pour mieux comprendre l'effectif cumulé croissant et décroissant.

Variables quantitatives continues

Effectifs cumulés

Dans le tableau suivant, on présente le nombre d'enfants de 1000 familles dans un village.

Nombre de d'enfants	1	2	3	4
Effectif	500	250	175	75

Variables quantitatives continues

Effectifs cumulés

l'effectif cumulé croissant de la première modalité est $N_1 = 500$.
l'effectif cumulé croissant de la deuxième est 750 ($N_2 = 500 + 250$).
l'effectif cumulé croissant de la troisième modalité est 925
($N_3 = 500 + 250 + 175$).
l'effectif cumulé croissant de "4" est 1000 ($N_4 = 500 + 250 + 175 + 75$).

De même on a détermine les effectifs cumulés décroissants de la façon suivante :

l'effectif cumulé croissant de la modalité 1 est 1000 ($\tilde{N}_1 = 1000 - 0$).
l'effectif cumulé croissant de la deuxième modalité est 500
($\tilde{N}_2 = 1000 - 500$).
l'effectif cumulé croissant de la troisième modalité est 250
($\tilde{N}_3 = 1000 - 500 - 250$).
l'effectif cumulé croissant de "4" est 75 ($\tilde{N}_4 = 1000 - 500 - 250 - 175$).

Variables quantitatives continues

Exemple

On résume alors les effectifs cumulés dans le tableau des effectifs croissants et décroissants suivant :

Tableau des effectifs cumulés

Nombre d'enfants	1	2	3	4
Effectifs cumulés croissant	500	750	925	1000
Effectifs cumulés décroissant	1000	500	250	75

Fréquences cas discret

Fréquences cumulées

Definition

On appelle fréquence cumulée croissante associée à la modalité x_k la somme des fréquences des modalités $x_1; \ldots; x_k$. c'est à dire

$$F_k = f_1 + \ldots + f_k = \sum_{i=1}^{k} f_i.$$

Fréquences cas discret

Fréquences cumulées

Ainsi, la fréquence cumulée croissante associée à la modalité x_k du caractère X est la fréquence d'individus de la population Ω pour lesquels la modalité correspondante x_k est inférieure ou égale à x_k.

Fréquences cas discret

Fréquences cumulées

Definition

On appelle fréquence cumulée décroissante de la valeur x_k le nombre

$$\tilde{F}_k = 1 - (f_1 + \ldots + f_{k-1}) = 1 - \sum_{i=1}^{k-1} f_i.$$

Fréquences cas discret

Fréquences cumulées

On considère l'exemple du nombre d'enfants des habitations d'un village. Ci-dessous le tableau statistique des fréquences correspondant à cette exemple.

Nombre d'enfants	1	2	3	4
Fréquence	0,5	0,25	0,175	0,075

Fréquences cas discret

Fréquences cumulées

La fréquence cumulée croissante de la première modalité est $F_1 = 0,5$.
La fréquence cumulée croissante de la deuxième modalité est
$F_2 =, 5 + 0,25 = 0,75$.
La fréquence cumulée croissante de la troisième modalité est
$F_3 = 0,5 + 0,25 + 0,175 = 0,925$.
La fréquence cumulée croissante de la quatrième modalité est
$F_4 = 0,5 + 0,25 + 0,175 + 0,075 = 1$.

Fréquences cas discret

Exemple

On fait la même chose pour les fréquences cumulés décroissantes

La fréquence cumulée décroissante de la première modalité est $\tilde{F}_1 = 1 - 0 = 1$.
La fréquence cumulée décroissante de la deuxième modalité est est $\tilde{F}_2 = 1 - 0,5 = 0,5$.
La fréquence cumulée décroissante de la troisième modalité est $\tilde{F}_3 = 1 - 0,5 - 0,25 = 0,25$.
La fréquence cumulée décroissante de la quatrième modalité est $\tilde{F}_4 = 1 - 0,5 - 0,25 - 0,175 = 0,075$.

Fréquences cas continue

Définition

Le calcul des effectifs et fréquences cumulés se fait de la même façon que pour le cas discret. En effet, la fréquence (ou effectif) cumulée d'une modalité x_k est la fréquence (ou effectif) des observations dont les modalités sont inférieures ou égales cette valeur x_k.

Cependant, dans le cas d'une variable continue dont les valeurs ont été regroupées sous forme de classes, c'est équivalent à chercher la fréquence d'observations qui sont strictement inférieures à la classe actuel. On considère l'exemple suivant pour avoir une idée du calcul des fréquences cumulées.

Fréquences cas continue

Exemple

Durée du trajet domicile-école

Tableau de distribution des fréquences

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Fréquence	0,6	0,3	0,1

Fréquences cas continue

Exemple

- La fréquence cumulée croissante de la première classe de $[5;\ 15[$ est $F_1 = 0,6$.
- La fréquence cumulée croissante de la deuxième classe $[5;\ 30[$ est égale à la fréquence de $[5;\ 15[$ plus la fréquence de $[15;\ 30[$, c-à-d $F_2 = 0,9$.
- La fréquence cumulée croissante de la dernière classe $[5;\ 40[$ est donc $F_3 = 1$, suivant le même principe.

Fréquences cas continue

Exemple

Classe	5	15	30	40
Fréquence cumulée croissante	0	0,6	0,9	1

Contents

Paramètres de position et de dispersion

Introduction

La moyenne, la médiane et le mode sont tous les trois des outils de base en statistiques connus sous le nom de **paramètres de position**.

les paramètres de dispersion indiquent de combien les valeurs d'une distribution s'écartent en général de la valeur de référence. Un paramètre de dispersion s'exprime toujours dans l'unité de mesure de la variable considérée. Les trois paramètres de dispersion les plus courants sont **l'étendue, l'intervalle interquartiles et l'écart type**.

Paramètres de position et de dispersion

Le Mode

Definition

Le Mode est un complément à la moyenne et à la médiane. Il permet de donner une indication statistique de tendance centrale à un ensemble de données. Le mode d'un ensemble d'observations est la valeur la plus fréquemment rencontrée.

On appelle mode d'une variable (qualitative ou quantitative discrète), la modalité ayant le plus grand effectif ou la plus grande fréquence. On note le mode par M_e.

Paramètres de position et de dispersion

Le Mode

Proposition

le Mode correspond aussi à (ou aux valeur(s)) ayant la plus grande hauteur dans la représentation graphique de la variable.

Paramètres de position et de dispersion

Le Mode

- Le Mode de la variable "nombre d'enfants" est $M_e = 1$
  car la fréquence de x_1 (50%) est la plus élevée, ou bien car l'effectif de x_1 ($n_1 = 500$) est le plus élevé.
- Le Mode de la variable "Matière préférée" est $M_e = "Chimie"$
  car la fréquence de la matière de chimie est ($f_1 = 40\%$) est la plus élevée, ou bien car l'effectif de la matière de chimie ($n_1 = 4$) est le plus élevé.

Paramètres de position et de dispersion

Le Mode

On passe maintenant au cas continue où on parle de classe modale.

Definition

Si les classes sont de même amplitude, on appelle classe modale d'une variable quantitative continue la (ou les classe(s)) ayant le plus grand effectif ou la plus grande fréquence.
C'est la classe correspondant au rectangle le plus haut dans l'histogramme des effectifs ou fréquences.

Paramètres de position et de dispersion

Le Mode

Durée du trajet domicile-école :

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Densité de fréquence	0.06	0.02	0.01

La classe modale de la variable "Durée du trajet domicile-école" est "$[5;\ 15[$".

Paramètres de position et de dispersion

Le Mode

Remarque :

- Il peut y avoir une ou plusieurs classes modales.
- Attention :
  Si les classes ne sont pas de même amplitude, il faut corriger les effectifs ou bien les fréquences. Donc comment corriger les effectifs.

Paramètres de position et de dispersion

Le Mode

Correction des effectifs :

Si les classes ne sont pas de même amplitude, on procède par correction de l'effectif en suivants ces étapes :

- Calculer les amplitudes $a_i = x_{i+1} - x_i$ de chaque classe.
- Choisir une amplitude de référence, en général celle qui se répète le plus. On la note a_r
- Corriger les effectifs (noté n'_i en utilisant la relation suivante :

$$n'_i = \frac{a_r \times n_i}{a_i},$$

où a_i est l'amplitude de la classe qu'on veut corriger son effectif.

Paramètres de position et de dispersion

Le Mode

Pour bien comprendre ce processus, on considère l'exemple suivant, où une statistique sur la durée du trajet entre domicile et la FSTBM de 10 étudiants :

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Effectif (n_i)	6	3	1
Amplitude (a_i)	10	15	10

Paramètres de position et de dispersion

Le Mode

On choisie comme classe de référence $[5;\ 15[$ puisque don amplitude se répète deux fois. Ensuite on corrige l'effectif de la deuxième classe $[15;\ 30[$ comme suit

$$n'_2 = \frac{10 \times 3}{15} = 2.$$

Donc le tableau avec l'effectif corrigé est donné comme suit :

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Effectif corrigé (n'_i)	6	**2**	1

La classe modale est donc $[5;\ 15[$ puisqu'elle a le plus grand effectif.

Paramètres de position et de dispersion

Le Mode

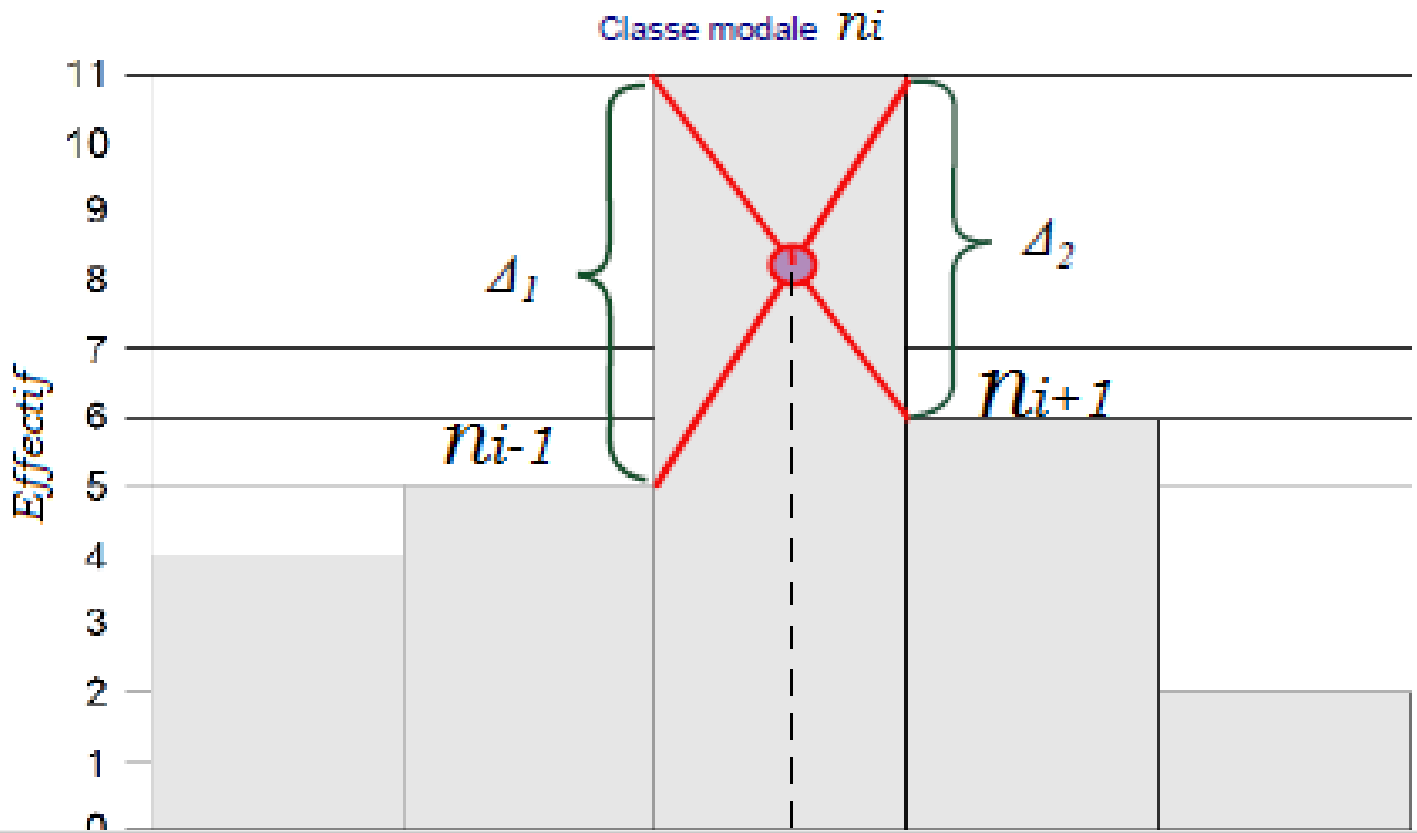
La question qui se pose maintenant est comment déterminer le Mode M_e ? En effet nous utilisons une interpolation linéaire qu'on peut la voir graphiquement.

Paramètres de position et de dispersion

Le Mode

Détermination du mode dans une classe :

Pour déterminer le mode à partir de la classe modale, on utilise une interpolation linéaire, inspiré de la Figure 12.



Paramètres de position et de dispersion

Le Mode

Le mode est donc déterminé par l'expression suivante :

$$M_e = \frac{\Delta_1}{\Delta_1 + \Delta_2}(x_{i+1} - x_i) + x_i. \tag{1}$$

avec :

$$\Delta_1 = n_i - n_{i-1},$$

et

$$\Delta_1 = n_i - n_{i+1}.$$

Si on revient maintenant à l'exemple précédent, on a $\Delta_1 = 6 - 0 = 6$ et $\Delta_2 = 6 - 2 = 4$, alors on a :

$$M_e = \frac{6}{6+4}(15 - 5) + 5 = 11.$$

Paramètres de position et de dispersion

La médiane

Soit S une série statistique. **Médiane :** la valeur centrale d'une série statistique dont les valeurs observées ont été rangées dans l'ordre croissant, est la valeur qui partage la population étudiée en deux sous-ensembles de même effectif. C'est la valeur de la variable statistique qui divise la population en deux sous-populations de même effectif. on la note par m. Comment alors calculer la médiane ?

Paramètres de position et de dispersion

La médiane (cas discret)

On distingue deux cas :

- Si l'effectif total N est impair, la médiane est la modalité du rang $\frac{N+1}{2}$, c'est :

$$m = x_{\frac{N+1}{2}}.$$

- Si l'effectif total N est pair, la médiane est la moyenne entre les deux variables de rang $\frac{N}{2}$ et $\frac{N}{2}+1$, c-à-d :

$$m = \frac{x_{\frac{N}{2}+1} + x_{\frac{N}{2}}}{2}.$$

Paramètres de position et de dispersion

La médiane (cas discret)

On considère l'exemple suivant pour illustrer le calcul de la médiane :

Soit un échantillon de 10 personnes dont l'âge est donné comme suit :

$$32,\ 18,\ 45,\ 38,\ 60,\ 64,\ 19,\ 20,\ 26,\ 24.$$

Il faut out d'abord arranger les valeurs de cette série par un ordre croissant, tel que :

$$18,\ 19,\ 20,\ 24,\ 26,\ 32,\ 38,\ 45,\ 60,\ 64.$$

Puisque l'effectif total $N = 10$ est pair, alors on applique directement la règle tel que :

$$m = \frac{x_{\frac{N}{2}+1} + x_{\frac{N}{2}}}{2} = \frac{26+32}{2} = 25.$$

Paramètres de position et de dispersion

La médiane (cas d'une variable continue)

Pour les données groupées e classe, il faut dans un premier temps définir la classe médiane (classe qui contient la médiane) à partie des effectif cumulés N_i. Ensuite, il faut envisager une interpolation linéaire autour de la médiane. On sait que $m \sim \frac{N}{2}$, donc il faut chercher la modalité de rang $\frac{N}{2}$ dans le tableau statistique. On considère l'exemple suivant pour mieux comprendre commet déterminer la médiane dans le cas continu.

Paramètres de position et de dispersion

La médiane (cas d'une variable continue)

On considère l'exemple qu'on a utilisé avant, où une statistique sur la durée du trajet entre domicile et l'ENSA de 10 étudiants :

Tableau de distribution des effectifs selon la durée du trajet

Durée	Effectif	Effectif cumulé
$[5;\ 15[$	3	3
$[15;\ 30[$	4	7
$[30;\ 40[$	3	10

Paramètres de position et de dispersion

La médiane (cas d'une variable continue)

On a $N = 10$, puisque $m \sim \frac{N}{2} = 5$. Il faut alors chercher la modalité du rang 5 dans le tableau des effectifs cumulés. Ceci correspond à la deuxième classe, donc $m \in [15;\ 30[$. Il reste à définir la médiane par interpolation. On a alors :

$$(1) \quad m \longrightarrow 5$$

$$(2) \quad 15 \longrightarrow 3$$

$$(3) \quad 30 \longrightarrow 7$$

On commence par l'opération suivante $(3) - (2)$, nous donne

$$(4) \quad 15 \longrightarrow 4$$

Paramètres de position et de dispersion

La médiane (cas d'une variable continue)

Après, on applique l'opération suivante $(1) - (2)$, ce qui donne :

$$(5) \quad m - 15 \longrightarrow 2.$$

On reprends maintenant les relations (4) et (5) et on utilise une règle à trois pour déterminer m, ce qui donne :

$$m - 15 = \frac{15 \times 2}{4}.$$

D'où :

$$m = 22.5$$

Paramètres de position et de dispersion

La moyenne

La moyenne, notée $\overline{x}$ est la moyenne arithmétique des modalités x_i pondérées par n_i. C'est un paramètre de position qui correspond au centre de gravité de la distribution.

Paramètres de position et de dispersion

La moyenne

Definition

Étant donné une série statistique quantitative discrète X prenant les valeurs $x_1; \ldots; x_p$ avec les effectifs partiels $n_1; \ldots; n_p$. Le nombre

$$\overline{x} = \frac{n_1 x_1 + \ldots + n_p x_p}{N} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i.$$

où $N = \sum_{i=1}^{p} n_i$ désigne l'effectif total de la série, s'appelle la moyenne arithmétique de la série X.

Paramètres de position et de dispersion

La moyenne

- La formule donnant $\overline{x}$ peut se réécrire $\overline{x} = \sum_{i=1}^{p} f_i x_i$ où $f_i = \frac{n_i}{N}$ est la fréquence de la valeur x_i.
- La moyenne de n nombres $x_1; \ldots; x_n$ est le quotient par n de la somme de ces nombres.

Paramètres de position et de dispersion

La moyenne

On considère l'exemple du nombre d'enfants de 1000 familles dans un village.

Tableau de distribution des effectifs

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

Paramètres de position et de dispersion

La moyenne

$$\overline{x} = \frac{500 \times 1 + 250 \times 2 + 175 \times 3 + 75 \times 4}{1000} = 1.825$$

Le nombre d'enfants moyen des habitations du village est $1,825 \sim 2$.

Remarque :

La moyenne n'est pas en général une valeur possible de la variable, il faut donc prévoir des arrondissements.

Paramètres de position et de dispersion

La moyenne

Tableau de distribution des fréquences

Nombre d'enfants	1	2	3	4
Pourcentage	50%	25%	17.5 %	7.5%

$$\overline{x} = 0.5 \times 1 + 0.25 \times 2 + 0.175 \times 3 + 0.075 \times 4 = 1.825$$

Paramètres de position et de dispersion

La moyenne (cas continu)

Pour une variable quantitative continue la moyenne arithmétique est aussi définie par l'une des formules précédentes en retenant pour x_i les centres des classes c_i.

Paramètres de position et de dispersion

La moyenne (cas continu)

De même que pour la moyenne arithmétique, dans le cas de variable quantitative continue, la variance peut être calculer à partir des données individuelles où à partir des données regroupées.

Paramètres de position et de dispersion

La moyenne (cas continu)

Durée du trajet domicile-école

Données individuelles :

$5 \square 6 \square 7 \square 10 \square 12 \square 13 \square 20 \square 25 \square 29 \square 39$

Données regroupées :

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Effectif	6	3	1
Fréquence	0.6	0.3	0.1
Centre	10	22.5	35

$$\overline{x} = \frac{6 \times 10 + 3 \times 22.5 + 1 \times 35}{10} = 16.25$$

La durée du trajet domicile-école moyenne est de 16.25 minutes

Paramètres de position et de dispersion

Remarques

- La moyenne calculée à partir des données individuelles est :

$$\overline{x} = \frac{5+6+7+10+12+13+20+25+29+39}{10} = 16.60$$

- La moyenne calculée sur les données regroupées n'est pas toujours égale à celle calculée sur les données individuelles. Problème d'arrondissement.
- La "vraie" valeur de la moyenne est celle calculée sur les données individuelles.
- La moyenne calculée sur les données regroupées est une valeur approchée de la bonne valeur de la moyenne ; il y a certainement une perte d'informations lors du regroupement des données en classes.

Paramètres de position et de dispersion

Autres moyennes

Definition

Soit r un entier rationnel non nul $(r \in \mathbb{Q}^*)$. On appel la moyenne d'ordre r de la série statistique $(x_i; n_i)_{1 \leq i \leq p}$ la quantité

$$\left[\frac{1}{N} \sum_{i=1}^{p} n_i x_i^r \right]^{\frac{1}{r}}$$

Paramètres de position et de dispersion

Autres moyennes

Definition

Selon la valeur de r on définit plusieurs moyennes

Paramètre	Moyenne	Formule
r = -1	harmonique	$\overline{x}_h = \frac{N}{\sum_{i=1}^{p} \frac{n_i}{x_i}}$
r $\approx$ 0	géométrique	$\overline{x}_g = \left[\prod_{i=1}^{p} x_i^{n_i}\right]^{\frac{1}{N}}$
r = +1	arithmétique	$\overline{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i$
r = +2	quadratique	$\overline{x}_q = \left[\frac{1}{N}\sum_{i=1}^{p} n_i x_i^2\right]^{\frac{1}{2}}$

Paramètres de position et de dispersion

Remarques

- La moyenne géométrique est un instrument permettant de calculer des taux moyens, notamment des taux moyens annuels. Son utilisation n'a un sens que si les valeurs ont un caractère multiplicatif.
- La moyenne quadratique est une moyenne qui trouve des applications lorsque l'on a affaire à des phénomène présentant un caractère sinusoïdal avec alternance de valeurs positives et de valeurs négatives. Elle est, de ce fait, très utilisée en électricité.
- On utilise la moyenne harmonique lorsqu'on veut déterminer un rapport moyen dans des domaines ou ils existent des liens de proportionnalité inverse.

Paramètres de position et de dispersion

Autres moyennes

definition

- La moyenne harmonique n'est définie que si toutes les observations sont non nulles. Cependant, la moyenne géométrique n'est calculable que si toutes les observations sont strictement positives.
- La comparaisons entre les différentes moyennes donne

$$\overline{x}_h \leq \overline{x}_g \leq \overline{x} \leq \overline{x}_q$$

Paramètres de position et de dispersion

L'étendue

Definition

On appelle étendue la différence entre la plus grande valeur et la plus petite valeur prise par la variable.

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

L'étendue vaut $4 - 1 = 3$.

Paramètres de position et de dispersion

L'étendue

L'étendue est une mesure de la dispersion des données. Il nous montre l'écart entre la petite et le grande valeur des modalités. On passe maintenant à un autre paramètre de dispersion plus représentatif.

Paramètres de position et de dispersion

La variance

Definition

Soit X une variable quantitative discrète définit sur une population Ω, d'effectif total N prenant les valeurs $x_1; \ldots; x_p$ avec les effectifs partiels $n_1; \ldots; n_p$. On appelle variance, notée σ^2 ou $V(X)$, de X le nombre

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2.$$

Paramètres de position et de dispersion

La variance

Tableau de distribution des effectifs

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

On a déjà calculé la moyenne $\overline{x} = 1,825$.

Nombre d'enfants (x_i)	1	2	3	4
$x_i - \overline{x}$	-0.825	0.175	1.175	2.175
$(x_i - \overline{x})^2$	0.681	0.031	1.381	4.731
Effectif (n_i)	500	250	175	75

$$\sigma^2 = \frac{(500 \times 0.681) + (250 \times 0.031) + (175 \times 1.381) + (75 \times 4.731)}{1000}$$
$$= 0.945$$

Paramètres de position et de dispersion

La variance

On montre que l'on peut encore écrire la variance sous la forme

$$\sigma^2 = \frac{\sum_{i=1}^{p} n_i x_i^2}{N} - \overline{x}^2 = \overline{x^2} - \overline{x}^2.$$

Avec cette relation, le calcul devient beaucoup plus rapide et plus pratique.

Paramètres de position et de dispersion

La variance

Démonstration :

Pour la démonstration, on commence par la définition de la variance :

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2.$$

Ensuite nous simplifions le terme $(x_i - \overline{x})^2$, on trouve alors :

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{p} n_i \left(x_i^2 + \overline{x}^2 - 2x_i\overline{x}\right) \\
&= \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 + \overline{x}^2 \frac{1}{N}\sum_{i=1}^{p} n_i - 2\overline{x}\frac{1}{N}\sum_{i=1}^{p} n_i x_i \\
&= \overline{x^2} + \overline{x}^2 - 2\overline{x}^2 \\
&= \overline{x^2} - \overline{x}^2.
\end{aligned}$$

Paramètres de position et de dispersion

La variance

Démonstration :

Car $\frac{1}{N}\sum_{i=1}^{p} n_i x_i = \overline{x}$ et $\frac{1}{N}\sum_{i=1}^{p} n_i = 1$.

Paramètres de position et de dispersion

La variance

Exemple :

On considère l'exemple suivant pour le calcul de la moyenne.

Nombre d'enfants (x_i)	1	2	3	4
x_i^2	1	4	9	16
Effectif (n_i)	500	250	175	75

$$\begin{aligned}\sigma^2 &= \frac{(500\times 1)+(250\times 4)+(175\times 9)+(75\times 16)}{1000} - (1.825)^2 \\ &= 0.945\end{aligned}$$

Paramètres de position et de dispersion

La variance

Remarque :

La variance peut s'écrire en utilisant les fréquences sous la forme suivante :

$$\sigma^2 = \sum_{i=1}^{p} f_i (x_i - \overline{x})^2$$

où $f_i = \frac{n_i}{N}$ est la fréquence de la valeur x_i.

De même on montre que

$$\sigma^2 = \sum_{i=1}^{p} f_i x_i^2 - \overline{x}^2$$

Paramètres de position et de dispersion

La variance

Tableau de distribution des proportions

Nombre d'enfants	1	2	3	4
Fréquence	0.5	0.25	0.175	0.075

On a déjà calculé la moyenne $\overline{x} = 1.825$.

Nombre d'enfants (x_i)	1	2	3	4
$x_i - \overline{x}$	-0.825	0.175	1.175	2.175
$(x_i - \overline{x})^2$	0.681	0.031	1.381	4.731
Fréquence (f_i)	0.5	0.25	0.175	0.075

$$\begin{aligned}\sigma^2 &= (0.5 \times 0.681) + (0.25 \times 0.031) + (0.175 \times 1.381) + (0.075 \times 4.731) \\ &= 0.945\end{aligned}$$

Paramètres de position et de dispersion

La variance

Le calcul est plus rapide et efficace, de plus, il y a moins risque d'erreurs par cette formule.

Nombre d'enfants (x_i)	1	2	3	4
x_i^2	1	4	9	16
Effectif (n_i)	0.5	0.25	0.175	0.075

$$\begin{aligned}\sigma^2 &= (0.5 \times 1) + (0.25 \times 4) + (0.175 \times 9) + (0.075 \times 16) - (1.825)^2 \\ &= 0.945\end{aligned}$$

Paramètres de position et de dispersion

La variance (cas continu)

Definition

Pour une variable quantitative continue la variance est aussi définie par les mêmes formules précédentes en remplaçant les x_i par les centres des classes c_i.

De même que pour la moyenne arithmétique, dans le cas de variable quantitative continue, la variance peut être calculer à partir des données individuelles où à partir des données regroupées.

Paramètres de position et de dispersion

La variance (cas continu)

- Calcul de la variance à partir des données individuelles :

$$\begin{aligned}\sigma^2 &= \frac{5^2+6^2+7^2+10^2+12^2+13^2+20^2+25^2+29^2+39^2}{10} - (1\\ &= 115.44\end{aligned}$$

- Calcul de la variance à partir des effectifs :

$$\begin{aligned}\sigma^2 &= \frac{(6\times 10^2)+(3\times 22.5^2)+(1\times 35^2)}{10} - (16.25)^2\\ &= 70.31\end{aligned}$$

- Calcul de la variance à partir des fréquences

$$\begin{aligned}\sigma^2 &= (0.6\times 10^2)+(0.3\times 22.5^2)+(0.1\times 35^2) - (16.25)^2\\ &= 70.31\end{aligned}$$

Paramètres de position et de dispersion

La moyenne (cas continu)

- La variance calculée à partir des données regroupées n'est pas toujours égale à celle calculée sur les données individuelles. Mais en général la différence n'est pas énorme c'est dû aux erreurs d'arrondissements.
- La valeur de la variance la plus précise est celle calculée sur les données individuelles.
- La variance calculée à parie des données regroupées est une valeur approchée de la valeur précise de la variance ; il y a eu certes une perte d'information lors du regroupement des données en classes.

Paramètres de position et de dispersion

L'écart-type

L'écart-type est une mesure de la dispersion d'une série statistique autour de sa moyenne. C'est le paramètre de dispersion le plus utilisé dans la littérature. On définit l'écart-type, noté σ, comme étant la racine carrée de la variance.

Paramètres de position et de dispersion

L'écart-type

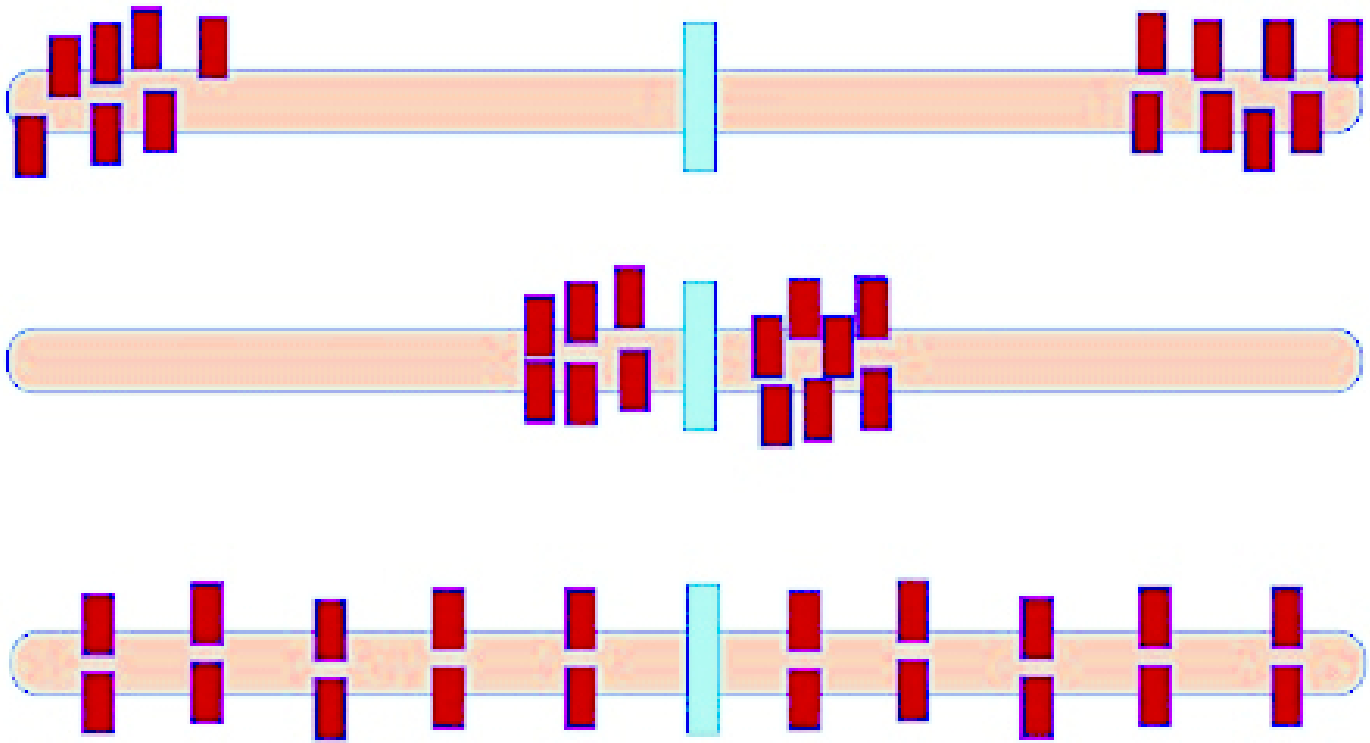


Figure – La dispersion d'une série statistique autour de sa moyenne.

Paramètres de position et de dispersion

L'écart-type

Le paramètre σ mesure la distance moyenne entre $\overline{x}$ et les valeurs de X. Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne. On dit alors que

- **Plus il est petit, plus les modalités sont concentrés autour de la moyenne** (on dit que la série est homogène).
- **Plus l'écart-type est grand, plus les données individuelles sont dispersées autour de la moyenne** (on dit que la série est hétérogène).

Paramètres de position et de dispersion

Coefficient de variation

Étant donné une variable quantitative X, de moyenne $\overline{x}$ et d'écart-type σ.
On appelle coefficient de variation de la variable X le rapport

$$C.V = \frac{\sigma}{\overline{x}}$$

Paramètres de position et de dispersion

Coefficient de variation

Le coefficient de variation mesure la dispersion relative des données individuelles par rapport à la moyenne. C'est un coefficient "sans dimension", invariant si on change l'unité de mesure.

Ce coefficient permet de comparer les dispersions de distributions qui ne sont pas exprimées dans la même unité (comme les salaires dans deux pays différents), ou deux distributions exprimées dans la même unité mais de moyennes différentes (comme les salaires dans deux entreprises du même pays).

Paramètres de position et de dispersion

Coefficient de variation

Les avantage de ce paramètre par rapport à l'écart-type :

- L'écart-type seul ne permet le plus souvent pas de juger de la dispersion des valeurs autour de la moyenne. Si par exemple une distribution a une moyenne de 10 et un écart-type de 1 (CV de 10 %), elle sera beaucoup plus dispersée qu'une distribution de moyenne 1000 et d'écart-type 10 (CV de 1 %).
- Ce nombre est sans unité, c'est une des raisons pour lesquelles il est parfois préféré à l'écart type qui lui ne l'est pas. En effet, pour comparer deux séries de données d'unités différentes, l'utilisation du coefficient de variation est plus judicieuse.

Paramètres de position et de dispersion

Les quartiles

En statistique descriptive, un quartile est chacune des trois valeurs qui divisent les données triées en quatre parts égales, de sorte que chaque partie représente 1/4 de l'échantillon de population. Le quartile fait partie des quantiles et il généralise la médiane.

Paramètres de position et de dispersion

Les quartiles

Le quartile est calculé en tant que 4-quartiles (voir Figure 14).

- Le 1er quartile est la donnée de la série qui sépare les 25% inférieurs des données (notation Q_1);
- Le 2e quartile est la donnée de la série qui sépare les 50% inférieurs des données (notation Q_2); il est également appelé médiane (noté aussi m); Le 3e quartile est la donnée de la série qui sépare les 75% inférieurs des données (notationQ_3); Par extension : le 0ème quartile est la donnée de la série qui sépare les 0% inférieurs des données (notation Q_0, c'est le minimum) et le 4ème quartile est la donnée de la série qui sépare les 0% supérieurs des inférieurs des données (notation Q_4, c'est le maximum)

Paramètres de position et de dispersion

Les quartiles

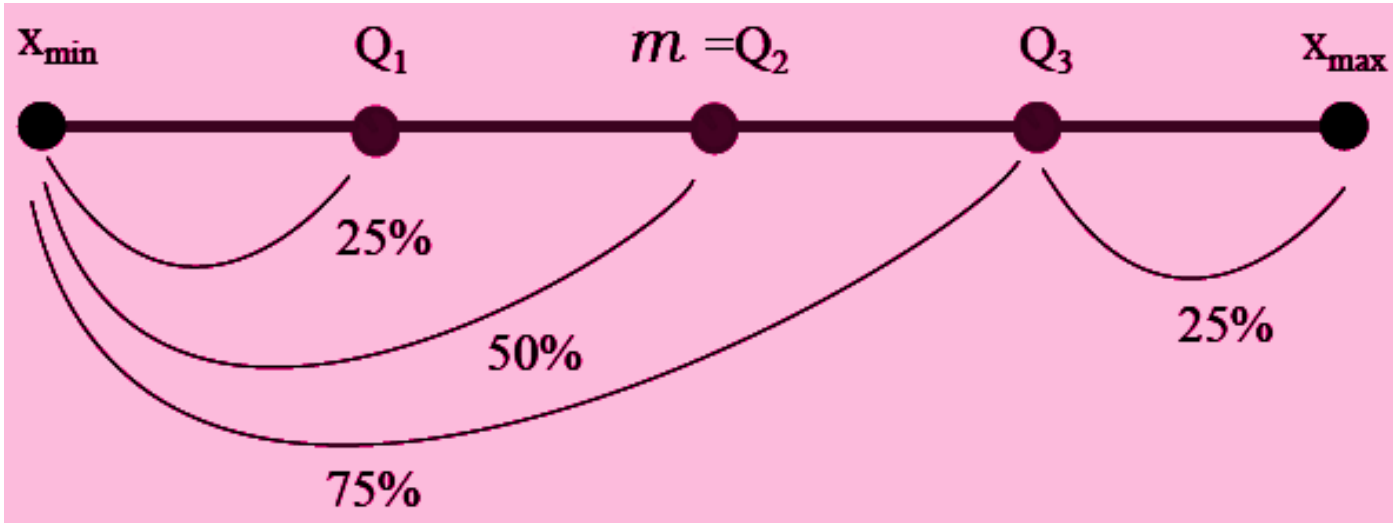


Figure – La répartition des quartiles.

Paramètres de position et de dispersion

Détermination des quartiles

La détermination des quartiles **dans le cas continu** se fait de la même façon que la médiane. C'est à partir de l'effectif cumulé du tableau statistique tel que :

$$Q_1 \longrightarrow \frac{N}{4}$$

$$Q_2 \longrightarrow \frac{N}{2}$$

$$Q_3 \longrightarrow \frac{3N}{4}.$$

Ensuite une interpolation linéaire est utilisée comme pour le cas de la médiane.

Paramètres de position et de dispersion

Détermination des quartiles

Dans le cas discret, on range les données par ordre croissant : si l'effectif total est N valeurs :

- Le quartile zéro (minimum) est celui qui a le rang 1.
- Le premier quartile est la modalité de rang $(N+3)/4$.
- La deuxième quartile (médiane) est celui qui a le rang $(2N+2)/4$ que l'on simplifie en (N+1)/2.
- Le troisième quartile est celui qui a le rang $(3N+1)/4$.
- le quatrième quartile est celui qui a le rang N.

Paramètres de position et de dispersion

Détermination des quartiles

L'âge de 13 personnes pris au hasard d'un village au Maroc est donné par ordre croissant comme suit :

$$4,\ 6,\ 13,\ 22,\ 28,\ 34,\ 40,\ 44,\ 50,\ 60,\ 62,\ 64,\ 70.$$

Calcul de Q_1 :
On applique la règle qui dit que $Q_1 \sim (N+3)/4 = \frac{16}{4} = 4$. Donc le premier quartile est la modalité du 4éme ordre, d'où :

$$Q_1 = 22.$$

Paramètres de position et de dispersion

Détermination des quartiles

Calcul de Q_3 :
On applique la règle qui dit que $Q_3 \sim (3N+1)/4 = \frac{3*13+1}{4} = 10$. Donc le troisième quartile est la modalité du 10éme ordre, d'où :

$$Q_3 = 60.$$

Calcul de Q_2 :
On applique la règle classique pour déterminer la médiane
$Q_2 \sim (N+1)/2 = \frac{13+1}{2} = 7$ (car N est impaire). Donc la médiane est la modalité du 7éme ordre, d'où :

$$Q_2 = 40.$$

Paramètres de position et de dispersion

Détermination des quartiles (cas d'une variable continue)

On considère l'exemple qu'on a utilisé avant, où une statistique sur la durée du trajet entre domicile et la FSTBM de 10 étudiants :

Tableau de distribution des effectifs selon la durée du trajet

Durée	Effectif	Effectif cumulé
$[5;\ 15[$	3	3
$[15;\ 30[$	4	7
$[30;\ 40[$	5	12

Paramètres de position et de dispersion

Détermination des quartiles (cas d'une variable continue)

On a $N = 12$, puisque $Q_1 \sim \frac{N}{4} = 3$. Il faut alors chercher la modalité du rang 3 dans le tableau des effectifs cumulés. Ceci correspond à la dernière classe, donc $Q_1 \in [5;\ 15[$. Il reste à définir Q_1 par interpolation. On a alors :

$$(1) \quad Q_1 \longrightarrow 3$$

$$(2) \quad 5 \longrightarrow 0$$

$$(3) \quad 15 \longrightarrow 3$$

On commence par l'opération suivante $(3) - (2)$, nous donne

$$(4) \quad 10 \longrightarrow 3$$

Paramètres de position et de dispersion

La médiane (cas d'une variable continue)

Après, on applique l'opération suivante $(1) - (2)$, ce qui donne :

$$(5) \quad Q_1 - 5 \longrightarrow 3.$$

On reprends maintenant les relations (4) et (5) et on utilise une règle à trois pour déterminer Q_1, ce qui donne :

$$Q_1 - 5 = \frac{10 \times 3}{3}.$$

D'où :

$$Q_1 = 15$$

Paramètres de position et de dispersion

Détermination des quartiles (cas d'une variable continue)

On a $N = 12$, puisque $Q_3 \sim \frac{3\times N}{4} = 9$. Il faut alors chercher la modalité du rang 9 dans le tableau des effectifs cumulés. Ceci correspond à la deuxième classe, donc $Q_3 \in [30;\ 40[$. Il reste à définir la médiane par interpolation. On a alors :

$$(1) \quad Q_3 \longrightarrow 9$$

$$(2) \quad 30 \longrightarrow 7$$

$$(3) \quad 40 \longrightarrow 12$$

On commence par l'opération suivante $(3) - (2)$, nous donne

$$(4) \quad 10 \longrightarrow 5$$

Paramètres de position et de dispersion

La médiane (cas d'une variable continue)

Après, on applique l'opération suivante $(1) - (2)$, ce qui donne :

$$(5) \quad Q_3 - 30 \longrightarrow 2.$$

On reprends maintenant les relations (4) et (5) et on utilise une règle à trois pour déterminer Q_3, ce qui donne :

$$Q_3 - 30 = \frac{10 \times 2}{5}.$$

D'où :

$$Q_3 = 34$$

Paramètres de position et de dispersion

Boîte à moustaches

Definition

La Boîte à moustaches, connue aussi sous le nom de Diagramme de Tukey ou box-plot, est un graphique représentatif, où on présente à la fois la médiane et les quartiles. Ceci permet de donner une idée sur l'asymétrie et la dispersion des valeurs ou modalités de la distribution statistique.

Paramètres de position et de dispersion

Boîte à moustaches

Exemple :

Durée du trajet domicile-école

Tableau de distribution des proportions

Classe	$[5;\ 15[$	$[15;\ 30[$	$[30;\ 40[$
Fréquence	6	3	1

Paramètres de position et de dispersion

Boîte à moustaches

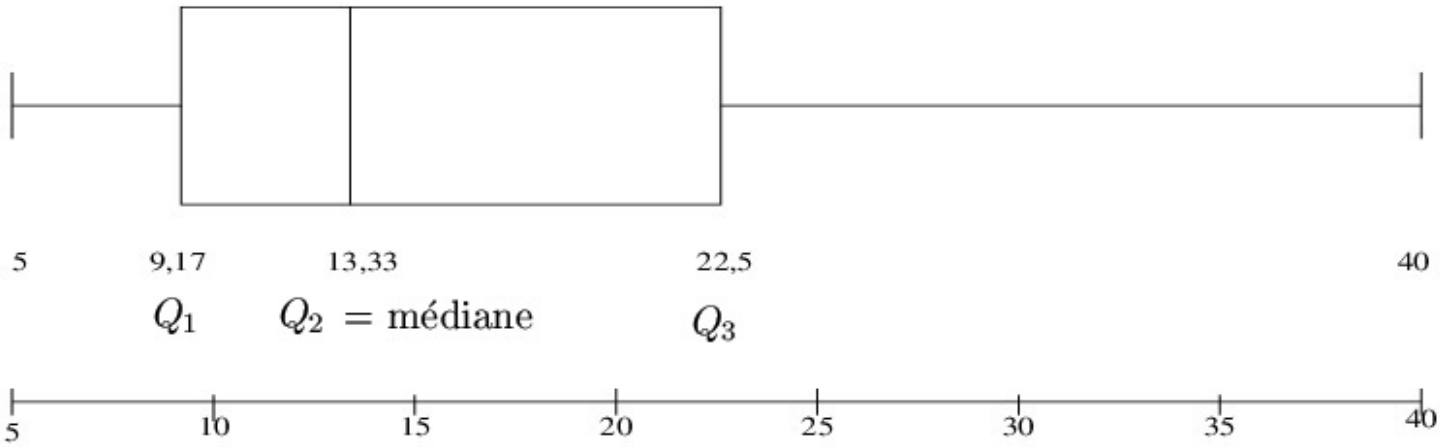
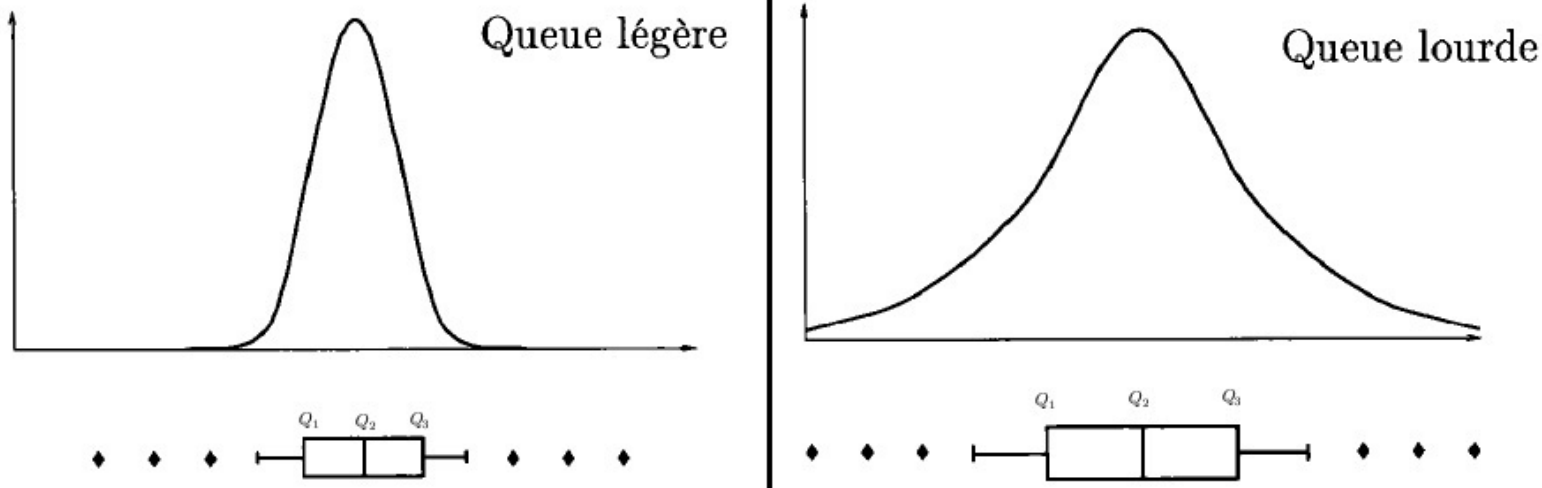


Figure – Le diagramme à moustache.

Paramètres de position et de dispersion

Boîte à moustaches

Pour maitre un lien entre le diagramme à moustache et la courbe statistique associée à la distribution étudiée, on considère ci-dessous une visualisation des caractéristiques d'une distribution à l'aide d'un box-plot.



Paramètres de position et de dispersion

Écart inter-quartiles

Definition

Un écart inter-quartiles est la différence $EI = Q_3 - Q_1$. Il mesure la dispersion des données autour de la médiane.

Exemple

Durée du trajet domicile-École

$[Q_1;\ Q_3] = [9,17;\ 22,5]$: 50% des élèves mettent entre 9,17 et 22,5 minutes pour aller à l'école.

25% mettent moins de 9,17 minutes et 25% mettent plus de 22,5 minutes.

Paramètres de position et de dispersion

Écart inter-quartiles

- L'intervalle inter-quartile est l'intervalle $[Q_1;\ Q_3]$.
  Il contient 50% des observations.
  25% des observations sont inférieures à Q_1 ; 25% des observations sont supérieures à Q_3.

Paramètres de position et de dispersion

Écart inter-quartiles

Remarques :

- Dans la boite de moustache on peut lire $Q_3 - Q_1$ comme taille de la moustache.
- L'écart inter-quartile est généralement utilisé pour détecter l'existence éventuelle des valeurs aberrantes.
  En effet, plusieurs logiciels statistiques comme le S-plus ou le SPSS, considèrent comme valeur aberrante, toute valeur x qui se trouve à l'extérieur de l'intervalle
  $\left[Q_1 - 1,5 \times (Q_3 - Q_1);\ Q_3 + 1,5 \times (Q_3 - Q_1)\right]$.

Contents

Paramètres de forme

Types de distributions

On distingue trois types de distributions selon leur symétrie par rapport à la moyenne comme la distribution normale, dissymétrique à gauche ou à droite. Il existe plusieurs paramètres de forme dans la littérature, nous allons traiter ici seulement quelques-uns.

Paramètres de forme

Moments centrés d'ordre r

Afin de distinguer les formes des distributions statistiques, on utilise des paramètres dites **moments centrés** de la variable statistique.

Definition

Soit X une variable quantitative définit sur une population Ω, d'effectif total N prenant les valeurs $x_1; \ldots; x_p$ avec les effectifs partiels $n_1; \ldots; n_p$ et m_r, le moment centré d'ordre $r \in \mathbb{Q}$ s'écrit

$$m_r = \frac{1}{N} \sum_{i=1}^{p} n_i (x_i - \overline{x})^r.$$

Paramètres de forme

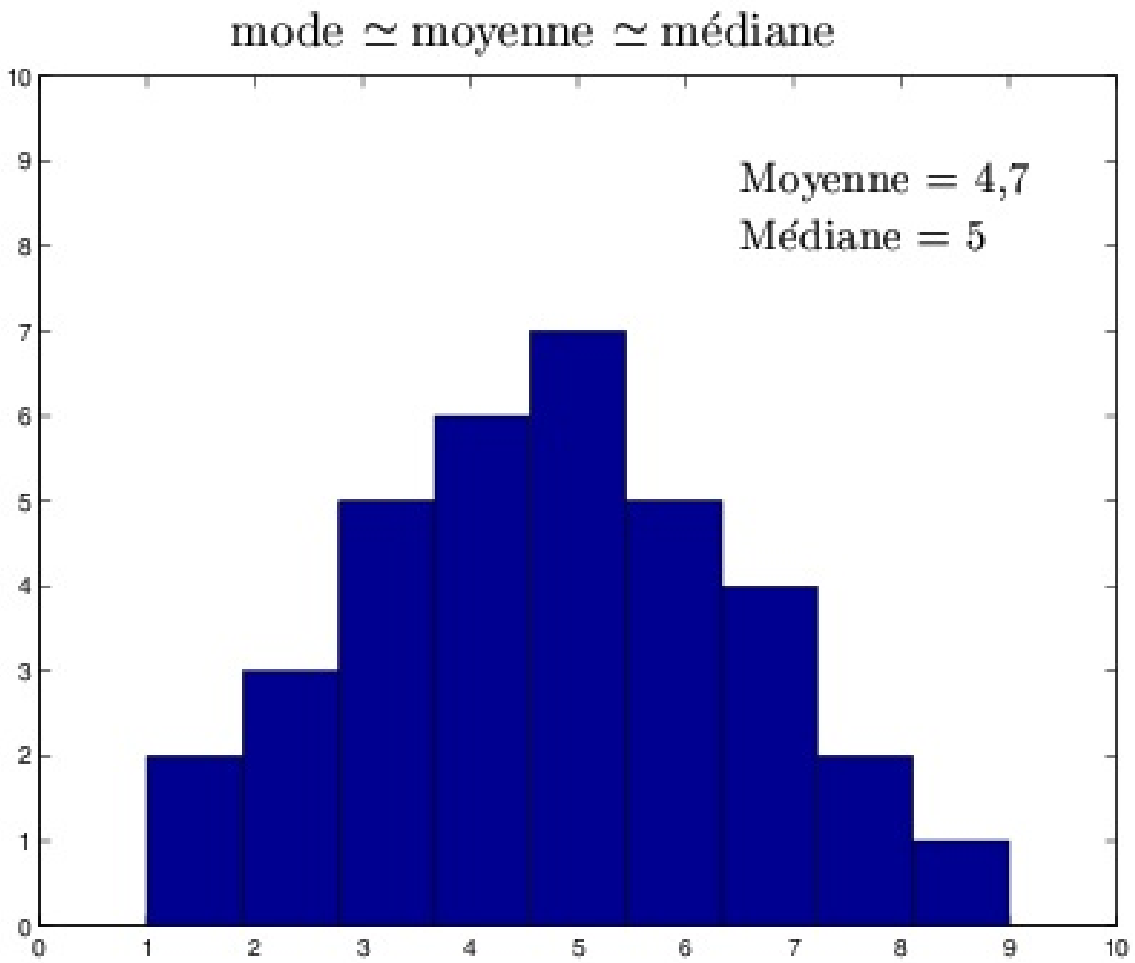
Moments centrés d'ordre r

Lorsque $r = 1$, le moment centré d'ordre 1 est nul : $m_1 = 0$, lorsque $r = 2$, le moment centré d'ordre 2 n'est rien d'autre que la variance.

Paramètres de forme

Comparaison Mode-Moyenne-Médiane

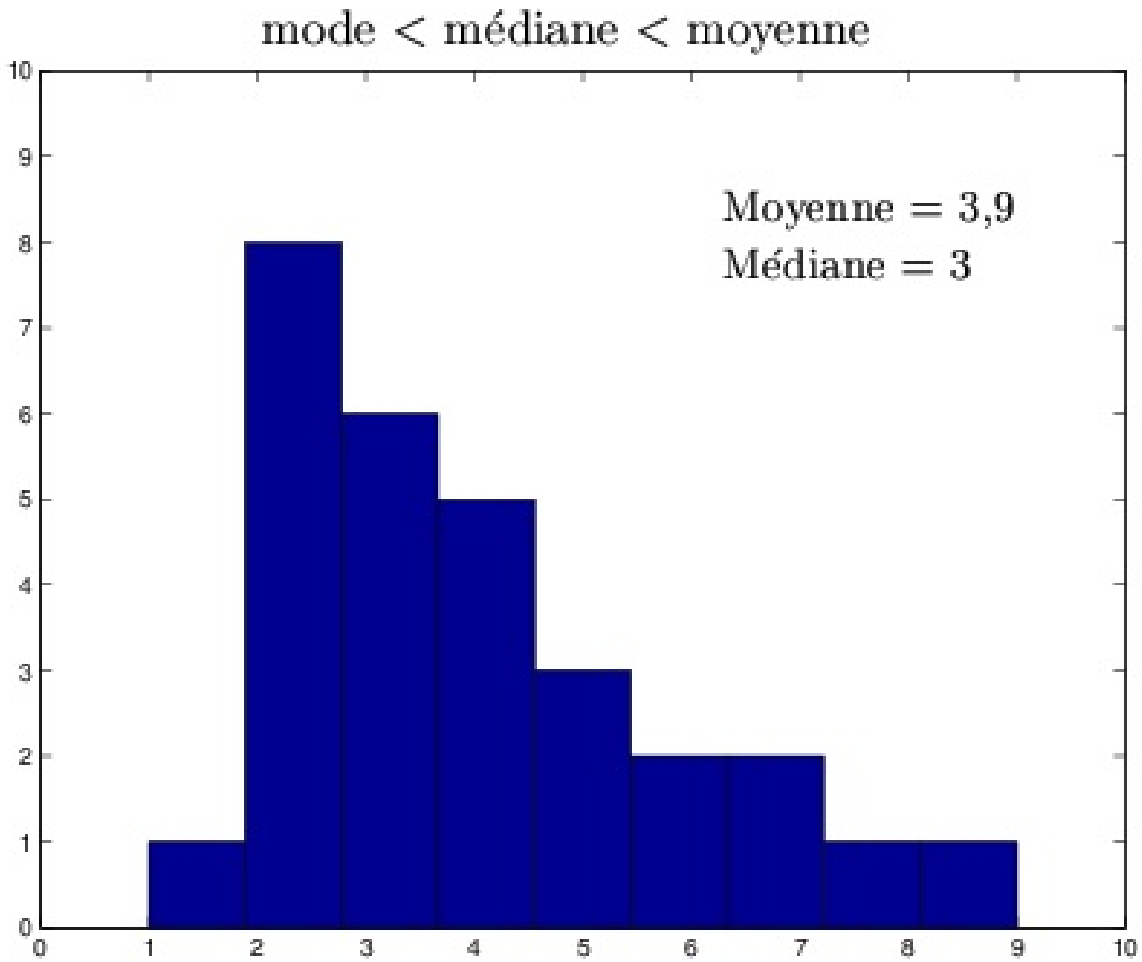
Si la distribution est symétrique



Paramètres de forme

Comparaison Mode-Moyenne-Médiane

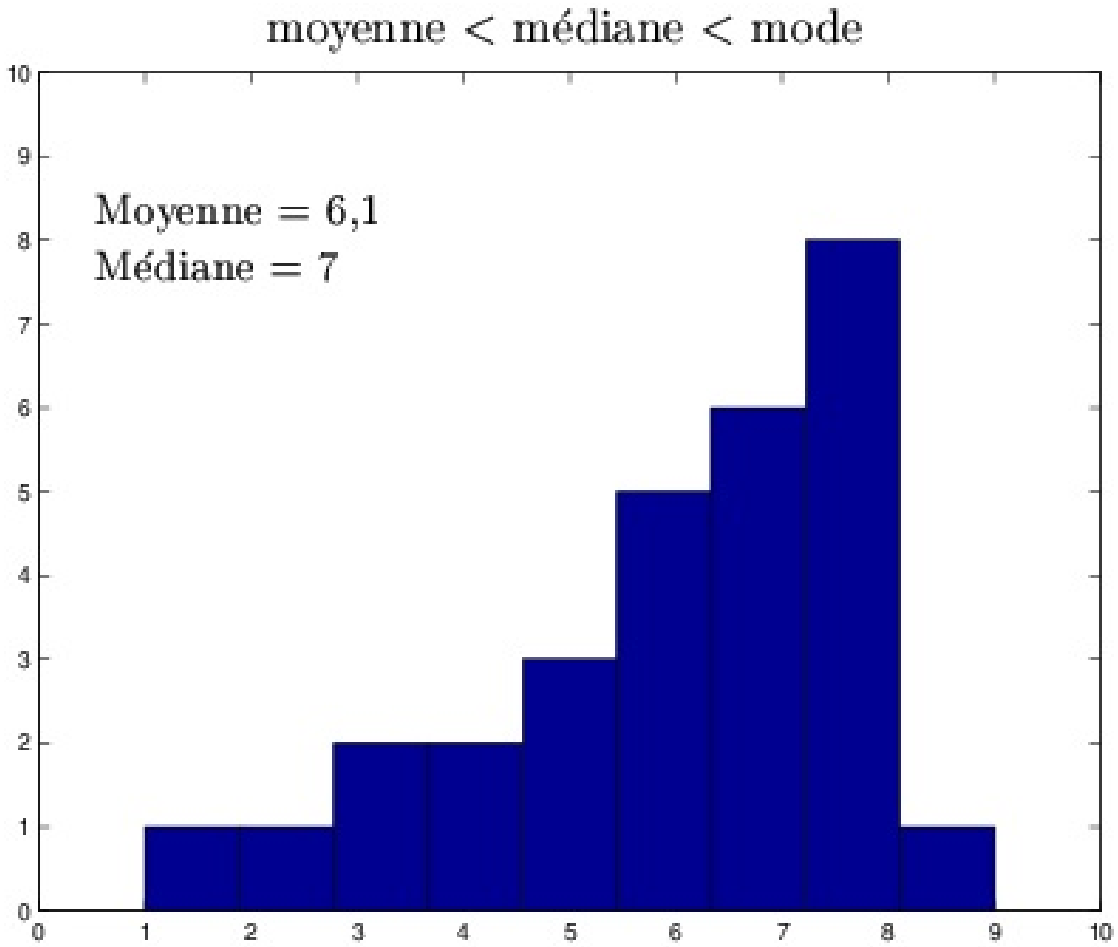
Si la distribution est dissymétrique étalée à droite :



Paramètres de forme

Comparaison Mode-Moyenne-Médiane

Si la distribution est dissymétrique étalée à gauche :



Paramètres de forme

Comparaison Mode-Moyenne-Médiane

Le paramètre le plus utilisé pour caractériser l'asymétrie d'une distribution $(x_i\ ;\ n_i)_i$ est le moment d'ordre 3.

On constate en effet que pour analyser la symétrie d'une distribution, il faut prévoir les cas suivants :

$$\begin{cases} \text{dissymétrique à gauche} & m_3 > 0 \\ \text{symétrique} & m_3 = 0 \\ \text{dissymétrique à droite} & m_3 < 0 \end{cases}$$

Paramètres de forme

Coefficient d'aplatissement

L'aplatissement d'une distribution est basé sur le moment centré d'ordre 4 définit comme suit :

$$m_4 = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{x})^4$$

Paramètres de forme

Coefficient d'aplatissement

Les paramètres d'aplatissement sont

- Le coefficient d'aplatissement de Pearson

$$\eta = \frac{m_4}{\sigma^4}$$

- Le coefficient d'aplatissement de Fisher

$$\beta = \frac{m_4}{\sigma^4} - 3$$

Plus la distribution observée est effilée, plus ces coefficients sont grands ; plus la distribution des données est aplatie, plus ces coefficients sont petits.

Paramètres de forme

Coefficient d'aplatissement de Pearson

Le coefficient de Pearson sert à mesurer l'aplatissement. Il s'agit d'un **coefficient sans dimension** avec les mêmes propriétés de Fisher.

Paramètres de forme

Coefficient d'aplatissement Fisher

Le calcul de ces coefficients pour la distribution normale donne $\eta = 3$ est par conséquent $\beta = 0$. Donc on prend l'aplatissement de la distribution normale comme aplatissement "référentiel" ainsi pour analyser l'aplatissement d'une distribution, on a les cas suivants :

$$\begin{cases} \text{Aplatie} & si \quad \beta < 0 \\ \text{Noramle} & si \quad \beta = 0 \\ \text{Effilée} & si \quad \beta > 0 \end{cases}$$

Contents

Statistique à deux dimensions

Introduction

La statistique à deux dimensions (ou analyses bivariée) consiste à étudier deux variables statistique d'une même population afin de déterminer une probable liaison, via des techniques descriptives ou probabilistes. Dans plusieurs phénomènes, on peut souvent envisager, pour une même population, **l'étude de deux ou plusieurs caractères en même temps**.

Statistique à deux dimensions

Exemples

- **Notes de Mathématique et notes de chimie** des étudiants de la filière BCG S3.
- **La taille et le poids** d'un échantillon de 50 personnes.
- **La durée de vie et le coût** d'une dizaine de matériaux.
- **L'âge et le chômage** d'une population d'une ville "A".
- **Nombre d'années d'ancienneté et revenus mensuels** des employés d'une entreprise.
- **Nombre des infectés du virus COVID19 et leur âge** au Maroc...

Statistique à deux dimensions

Objectifs

Il y a plusieurs objectifs pour réaliser ces études tels que :

- **Étendre les notions de la statistique descriptive** à une variable au cas de deux variables.
- **La mise en évidence d'un lien ou d'une absence de lien entre les deux variables** étudiées.
- **Caractériser et commenter les relations qui peuvent exister entre deux séries** d'observations considérées simultanément (sens, intensité).

Statistique à deux dimensions

Notations

Dans la suite, on considère une population P de N individus pour laquelle on présente **deux observations relatives à deux caractères (variables statistiques)** X et Y pouvant être de nature différente (qualitatif, quantitatif discret ou continu).

Statistique à deux dimensions

Notations

Comme on a vu en statistique descriptive à une variable, on envisagera trois aspects pour présenter et analyser les données statistiques relatives à un couple de caractères $(X\,;\,Y)$:

- ⋆ **Tableaux statistiques** (à deux variables),
- ⋆ **Représentations graphiques** (à deux variables),
- ⋆ **Les paramètres indiquant la relation entre les variables étudiés**.

Statistique à deux dimensions

Notations

On commence par le tableau statistique qui caractérise la présentation des deux variables en même temps. On peut alors présenter les données observées dans une population relatives à deux variables X et Y, considérées simultanément, sous forme d'une distribution d'effectifs ou de fréquences à deux dimensions, appelées distribution conjointe du couple (X, Y). Pour cela on a recours à un tableau statistique, appelé **Tableau de contingence**.

Statistique à deux dimensions

Tableau de contingence

Soient X et Y deux variables statistiques qualitatives ou quantitatives définies sur la même population finie Ω. Le couple (X, Y) **s'appelle une distribution statistique conjointe**. Soient $X(\Omega) = \{x_1, \ldots, x_r\}$ et $Y(\Omega) = \{y_1, \ldots, y_s\}$ l'ensemble des modalités ou valeurs prises par chacune des variables X et Y.

Statistique à deux dimensions

Tableau de contingence

Definition

On appelle effectif partiel n_{ij} du couple $(x_i; y_j)$ le nombre des individus de la population Ω pour lesquels le caractère X prend la valeur x_i et le caractère Y prend la valeur y_j. C-à-d, Chaque case (ligne i et colonne j) du tableau indique l'effectif n_{ij} des individus présentant la modalité (x_i ; y_j) du couple (X ; Y).

Statistique à deux dimensions

Tableau de contingence

Le tableau de contingence est donc un tableau à deux dimensions (double entrée) présenté sous la forme suivante : (Table??).

Table – Tableau de contingence

X\ Y	y_1	y_2	...	y_j	...	y_s	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	N

Statistique à deux dimensions

Tableau de contingence

Dans ce tableau :

- i désigne l'indice d'une ligne et j désigne l'indice d'une colonne.
- n_{ij} désigne l'effectif partiel.
- Effectif partiels marginaux :
  On note $n_{i.}$ **l'effectif marginal de** X (effectif total en lignes)

$$n_{i.} = \sum_{j=1}^{s} n_{ij}$$

On note $n_{.j}$ **l'effectif marginal de** Y (effectif total en colonnes)

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

Statistique à deux dimensions

Tableau de contingence

Si n_{ij} est l'effectif partiel d'un couple de valeurs $(x_i; y_j)$, et N l'effectif total de la population, le rapport

$$f_{ij} = \frac{n_{ij}}{N},$$

s'appelle **la fréquence partielle** du couple $(x_i; y_j)$.

Statistique à deux dimensions

Tableau de contingence

- On appelle **fréquence marginale $f_{.j}$ de la valeur** y_j de la variable Y le rapport

$$f_{.j} = \frac{n_{.j}}{N}$$

- On appelle **fréquence marginale $f_{i.}$ de la valeur** x_i de la variable X le rapport

$$f_{i.} = \frac{n_{i.}}{N}$$

Statistique à deux dimensions

Tableau de contingence

Definition

On appelle distribution conjointe des effectifs (ou fréquences) d'une série statistique double (X, Y) définie sur une population Ω, l'application définie par :

$$f : X(\Omega) \times Y(\Omega) \mapsto \mathbb{N}$$

$$(x_i; y_j) \longrightarrow n_{ij} \text{ ou } f_{ij}$$

chaque couple de valeur $(x_i; y_j)$, associé à son effectif n_{ij} (ou à sa fréquence f_{ij}). C'est **l'intérieur du tableau de contingence**.

Statistique à deux dimensions

Tableau de contingence

Definition

On appelle distribution marginale de X la distribution à une dimension des observations relatives au caractère X indépendamment des observations relatives au caractère de Y. Elle correspond aux effectifs $n_{i.}$: (ou aux fréquence $f_{i.}$:) de **la dernière colonne du tableau de contingence**.
De la même façon, on définit **la distribution marginale de Y**. Elle correspond aux effectifs $n_{.j}$ (ou aux fréquence $f_{.j}$) de **la dernière ligne du tableau de contingence**.

Statistique à deux dimensions

Exemple

Dans une entreprise de 200 salariés, on étudie les variables X : âge et Y :

	X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	Total
salaires.	[20,22[(i=1)	14	6	20
	[22,24[(i=2)	28	46	74
	[24,26[(i=3)	20	86	106
	Total	62	138	200

Statistique à deux dimensions

Exemple

X et Y sont des variables continues (regroupées en classes).
On note I le nombre de modalités de X (ici I = 3) et J le nombre de modalités de Y (ici J = 2).
$n_{12} = 6$ salariés sont âgés entre 20 et 22 ans et ont un salaire compris entre 1000 et 1200.
$n_{2.} = 74$ salariés sont âgés entre 22 et 24 ans ; $n_{.1} = 62$ salariés ont un salaire ente 800 et 1000.
Tableau des Fréquences partielles et marginales :

Statistique à deux dimensions

Exemple

X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	La distribution marginale de X
[20,22[(i=1)	7%	3%	10%
[22,24[(i=2)	14%	23%	37%
[24,26[(i=3)	10%	43%	53%
La distribution marginale de Y	31%	69%	100 %

Statistique à deux dimensions

Distributions conditionnelles

On appelle **distribution conditionnelle de X sous la condition** $Y = y_j$, la distribution à une dimension des individus de **la population relativement au caractère X sachant que la modalité de Y est** y_j . **Elle correspond aux effectifs** n_{ij} **de la colonne j du tableau de contingence**. De façon analogue, on définit la distribution conditionnelle de Y sous la condition $X = x_i$. Elle correspond aux effectifs n_{ij} de la ligne i du tableau de contingence.

Une **distribution conditionnelle** est une distribution statistique soumise à une condition sur la population à un événement particulier (une classe par exemple).

Statistique à deux dimensions

Distributions conditionnelles

On revient alors à l'exemple précédent : On a le nombre de colonne est deux (J = 2) $\Rightarrow$ il y a deux distributions conditionnelles de X par rapport à Y. On a par exemple :

1. la distribution de X sachant $Y \in [800; 1000[$.
2. la distribution de X sachant $Y \in [1000; 1200[$.

Statistique à deux dimensions

Distributions conditionnelles

Pour la distribution de Y sous la condition $X = x_i$, on a trois lignes (I = 3) $\Rightarrow$ il y a alors trois distributions conditionnelles de Y par rapport à $X = x_i$, à savoir :

1. la distribution de Y sachant $X \in [20; 22[$
2. la distribution de Y sachant $X \in [22; 24[$
3. la distribution de Y sachant $X \in [24; 26[$

Statistique à deux dimensions

Fréquences conditionnelles de X sachant Y

La fréquence conditionnelle de la modalité x_i du caractère X sachant que $Y = y_j$, notée f_{x_i/y_j} , est la proportion d'individus présentant la modalité x_i du caractère X par rapport au total des individus présentant la modalité y_j du caractère Y :

$$f_{xi/yj} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}.$$

On présentera alors ces fréquence sur le tableau de contingence sous forme de pourcentage, exemple :

Statistique à deux dimensions

Fréquences conditionnelles de X sachant Y

X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	Total
[20,22[(i=1)	22.6%	4.3%	.
[22,24[(i=2)	45.2%	33.3%	.
[24,26[(i=3)	32.2%	62.4%	.
Total	100%	100%	.

Statistique à deux dimensions

Fréquences conditionnelles de X sachant Y

On calcule les fréquences des âges en considérant la condition sur le salaire des individus entre 800 et 1000, puis à la sous-population des individus ayant un salaire entre 1000 et 1200.

Interprétation :

- 22.6%$= \frac{n_{11}}{n_{.1}} = \frac{7}{31}$ des employés ayant un salaire entre 800 et 1000 sont âgés entre 20 et 22 ans.
- Parmi les employés ayant un salaire entre 1000 et 1200, 62.4%$= \frac{n_{32}}{n_{.2}} = \frac{86}{138}$ d'entre eux sont âgés entre 24 et 26 ans.

Statistique à deux dimensions

Fréquences conditionnelles de X sachant Y

La fréquence conditionnelle de la modalité y_j du caractère Y sachant que $X = x_i$, notée f_{y_j/x_i} , est la proportion d'individus présentant la modalité y_j du caractère Y par rapport au total des individus présentant la modalité x_i du caractère X :

$$f_{y_j/x_i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}.$$

On présentera alors ces fréquences sur le tableau de contingence sous forme de pourcentage, exemple :

Statistique à deux dimensions

Fréquences conditionnelles de X sachant Y

X\ Y	[800,1000[(j=1)	[1000,1200[(j=2)	Total
[20,22[(i=1)	70%	30%	100%
[22,24[(i=2)	37.8%	62.2%	100%
[24,26[(i=3)	18.9%	81.1%	100%
Total	.	.	.

Statistique à deux dimensions

Fréquences conditionnelles de X sachant Y

Interprétation :

- 70% des employés âgés entre 20 et 22 ans ont un salaire compris entre 800 et 1000.
- Parmi les employés âgés entre 22 et 24 ans, 62.2% d'entre eux ont un salaire compris entre 1000 et 1200.

Statistique à deux dimensions

Représentation graphique

Sur les distributions partielles (marginales et conditionnelles), on peut tracer des graphes (diagramme en tuyau d'orgue, diagramme en bâton, histogramme, fréquences cumulées) et on peut calculer des paramètres statistiques de la même façon que pour les distributions d'une seule variable (vus dans le chapitre précédant). Alors que pour représenter graphiquement une distribution conjointe de deux variables qualitatives, plusieurs possibilités sont envisageables. On peut citer par exemple :

- Histogramme groupé
- Histogramme empilé

Statistique à deux dimensions

Représentation graphique

Pour le histogramme groupé, son principe est simple, on fixe une variable sur l'axe des abscisses et on trace des rectangles dont la hauteur est proportionnelle au effectif (ou fréquence) de l'autre variable.

Statistique à deux dimensions

Représentation graphique

On traite la relation entre les distributions des observations de Maths et de Chimie des étudiants de BCG en deuxième année à la FST de Béni Mellal. On traite alors deux caractères qualitatifs ordinales. Les modalités sont alors comme suit : Médiocre (M), passable (P), bien (B), très bien (TB) pour le module de Physique et NV, R, validé (V) pour le module de Math. On remarque que les modalités sot présentés de façon différente. Les données sont présentées comme suit :

Maths\ Chimie	NV (j=1)	R (j=2)	V (j=3)
M (i=1)	10	20	5
P (i=2)	5	10	15
B (i=3)	5	3	10
TB (i=4)	20	7	10

Statistique à deux dimensions

Représentation graphique

Le histogramme groupé des effectif associé à cette distribution est donné par :

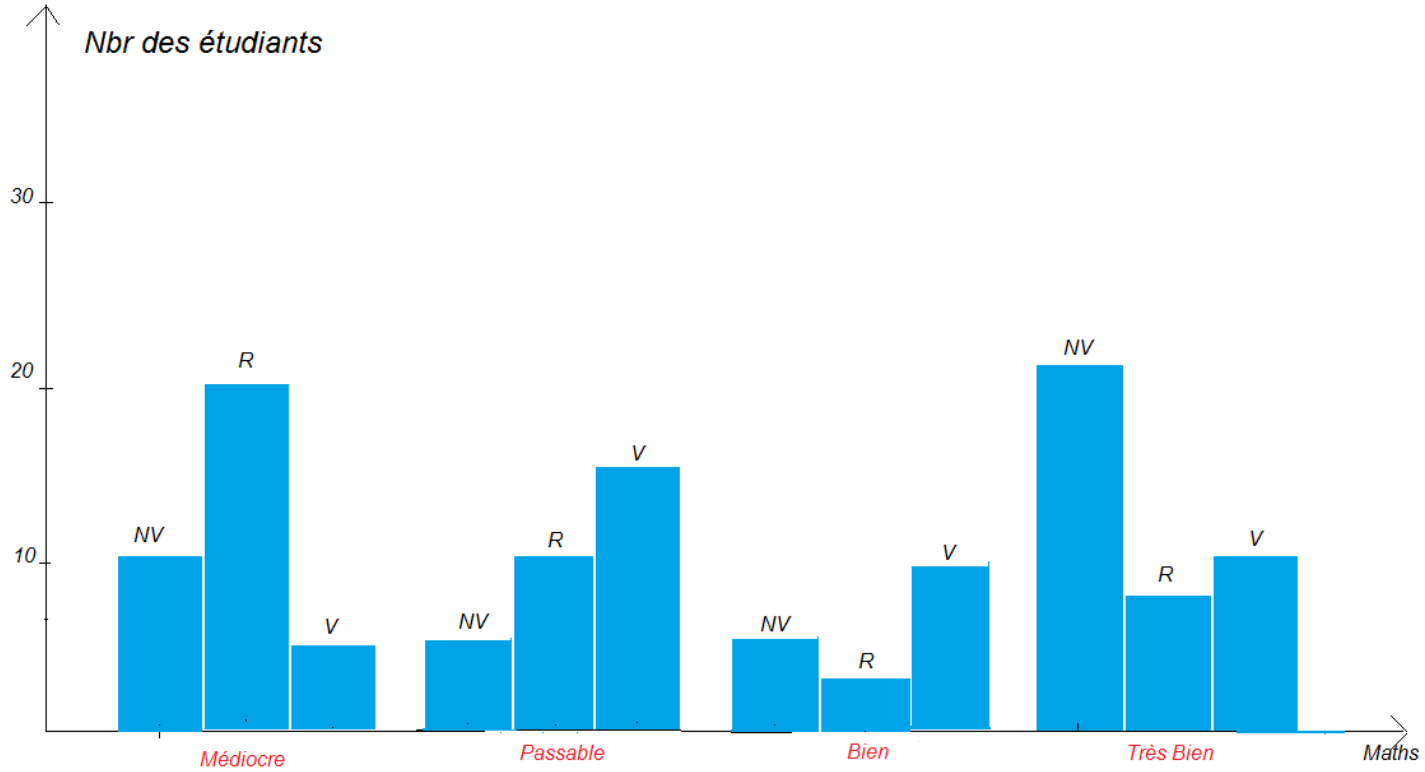


Figure – Le histogramme groupé associé aux observations des notes de Maths en

Statistique à deux dimensions

Représentation graphique

Pour les caractères de type quantitatif, on envisagera une étude de paramètre. En effet, on distingue deux types de paramètres :

- Les paramètres qui concernent une seule variable, ils caractérisent les distributions partielles (marginales et conditionnelles).
- Les paramètres qui décrivent les relations qui existent entre les deux variables considérées simultanément, ils caractérisent la distribution conjointe.

Dans la suite, nous nous intéresserons aux paramètres décrivant la relation entre deux variables, plus précisément, nous nous focaliserons sur la covariance.

Statistique à deux dimensions

Covariance

Si (X, Y) désigne un couple de série statistiques quantitatives définies sur une même population Ω, prenant respectivement les valeurs $x_1, \ldots, x_r$ et $y_1, \ldots, y_s$, et si n_{ij} désigne l'effectif partiel du couple $(x_i; y_j)$, on appelle covariance du couple (X, Y) et on note $Cov(X, Y)$ le nombre

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}(x_i - \overline{X})(y_j - \overline{Y})$$

avec $\overline{X}$ est la moyenne de X et $\overline{Y}$ est la moyenne de Y. La covariance peut s'écrire d'une autre façon plus simple dans le calcul : ou bien

$$Cov(X, Y) = \sum_{i=1}^{r} \sum_{j=1}^{s} f_{ij}(x_i - \overline{X})(y_j - \overline{Y})$$

Statistique à deux dimensions

Covariance

On présente ici quelque propriétés intéressantes de la covariance :

- $Cov(X,X) = \sigma_X^2$ (la variance de X)
- $Cov(Y,Y) = \sigma_Y^2$ (la variance de Y)
- $Cov(X,Y) = Cov(Y,X)$ (la covariance est symétrique)
- $Cov(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_iy_j - \overline{XY}$ (en utilisant les fréquences.
- Si X et Y sont indépendantes alors $Cov(X,Y) = 0$

Statistique à deux dimensions

Covariance

La moyenne et la variance de la variable X est donnée par les formules suivantes :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{r} n_{i.}x_i = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}x_i$$

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{r} n_{i.}(x_i - \overline{X})^2 = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - \overline{X})^2$$

Statistique à deux dimensions

Corrélation

On dit qu'il y a corrélation entre deux variables lorsqu'elles ont tendance à varier soit toujours dans le même sens (X augmente, Y a tendance à augmenter aussi), soit en sens inverse (X augmente, Y a tendance à diminuer).
Des questions se posent :

- Comment peut-on quantifier cette liaison ?
- Comment peut-on tester si cette liaison est statistiquement significative ?

Pour mesurer l'intensité de la relation entre deux variables on définit le coefficient de corrélation de Pearson , connu par son efficacité et sa précision.

Statistique à deux dimensions

Coéfficient de corrélation linéaire

Le coefficient de corrélation de Pearson ρ est une mesure du degré d'association linéaire entre deux variables quantitatives X et Y.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

ρ est un nombre compris entre -1 et 1.

Statistique à deux dimensions

Coéfficient de corrélation linéaire

Remarques :

- Si $\rho > 0$ cela signifie que **les deux variables évoluent en même sens (sens positif)**.
- Si $\rho < 0$ cela signifie que **les deux variables évoluent en sens contraire (sens négatif)**.
- Lorsque $\rho = 0$, on dit que X et Y sont non corrélés : **il n'y a pas d'association linéaire** entre X et Y.
- Si $\rho \pm 1$ alors l'une des variable est une fonction affine de l'autre, (Y est une fonction affine de X i.e. $Y = aX + b$ avec b du signe de ρ).

Statistique à deux dimensions

Coéfficient de corrélation linéaire

Remarque :
Lorsque X et Y sont indépendantes $\rho = 0$ mais la réciproque est en général fausse.

Statistique à deux dimensions

Régression linéaire

La régression linéaire permet d'étudier la relation entre deux variables quantitatives, en étudiant les variations de l'une en fonction des valeurs de l'autre. Dans le cas où tous les points $(x_i\,;\,y_i)$ du diagramme de régression sont alignés, on dit que la relation entre X et Y est une dépendance fonctionnelle linéaire.

Statistique à deux dimensions

Régression linéaire

Si les points du diagramme de régression ne sont pas tous alignés, on doit mesurer le degré de dépendance linéaire entre les deux variables. Lorsque l'on a estimé la droite de régression, on doit se demander si cette estimation est de bonne qualité.

Statistique à deux dimensions

Régression linéaire

Pour cela, on envisage un nouveau tableau de tel sorte que chaque couple de variables est sans pondération. On présente alors les données sous la forme suivante :

Modalités de X	x_1	x_2	x_3	$\ldots$	x_i	$\ldots$	x_N
Modalités de Y	y_1	y_2	y_3	$\ldots$	y_i	$\ldots$	y_N

Statistique à deux dimensions

Régression linéaire

Donc pour réaliser une bonne analyse de la relation entre ces deux variables, on se pose les questions suivantes :

- ⋆ Quelle est la nature de la relation qui existe entre les deux variables X et Y ?
- ⋆ Si on connaît la valeur de l'une des variables, peut-on estimer la valeur de la deuxième variable ?

$$Y = f(X)$$

- ⋆ Comment mesurer la précision de l'estimation ?

Statistique à deux dimensions

Régression linéaire

Pour répondre à ces questions, on commence par représenter le nuage de points $Mi(x_i; y_i)$ dans un repère cartésien. Ensuite, on cherchera s'il existe une droite ou une courbe qui passe par le plus possibles des points M_i pour qu'elle soit une bonne approximation du nuage de points. Ceci indiquera s'il y a une dépendance linéaire entre les deux variables. Si c'est le cas, les points du nuage seront concentrés autour de la droite de la fonction, et il devient possible de prévoir la seconde variable si on connaît la première. Ce nuage de points avec la courbe d'ajustement linéaire est appelé **Diagramme de régression**.

Statistique à deux dimensions

Régression linéaire

La solution de ce problème est particulièrement simple lorsqu'on cherche une fonction linéaire

$$Y = aX + b$$

Le critère utilisé pour optimiser l'approximation est le critère des moindres carrés. On parle alors de régression (ou ajustement) linéaire de Y en fonction de X. Mais il est possible d'avoir d'autres types de régression non linéaires : exponentielle, logarithmique, polynômiale, etc. En prenant compte des erreurs entre les points, le modèle linéaire s'écrit

$$Y = aX + b + \epsilon,$$

où ϵ est une variable aléatoire appelée erreur résiduelle.

Statistique à deux dimensions

Régression linéaire

On considère par exemple deux séries statistiques quantitatives, X représentant les notes de Maths et Y, représentant les notes de Chimie sur la même population Ω. Soit $X(\Omega) = \{x_1, \ldots, x_10\}$ l'ensemble des valeurs prises par X. $Y(\Omega) = \{y_1, \ldots, y_10\}$ l'ensemble des valeurs prises par Y.

Statistique à deux dimensions

Régression linéaire

Les données sont réparties et **arrangées dans un ordre croissant** dans le tableau de pondération suivant :

Étudiant N	1	2	3	4	5	6	7	8	9	10
Note de Maths (X)	3	6	7	9	10	12	13	15	17	19
Note de Chimie (Y	2	4.5	5.5	7	8	10	12.5	14	16	17

Statistique à deux dimensions

Régression linéaire

On remarque d'après ces données que plus la note de Maths est élevée, celle de Chimie l'est aussi et inversement proportionnel. Il y donc certainement une corrélation dans le sens positive entre ces deux variables. On obtient le nuage de points suivant qui représente le Diagramme de régression :

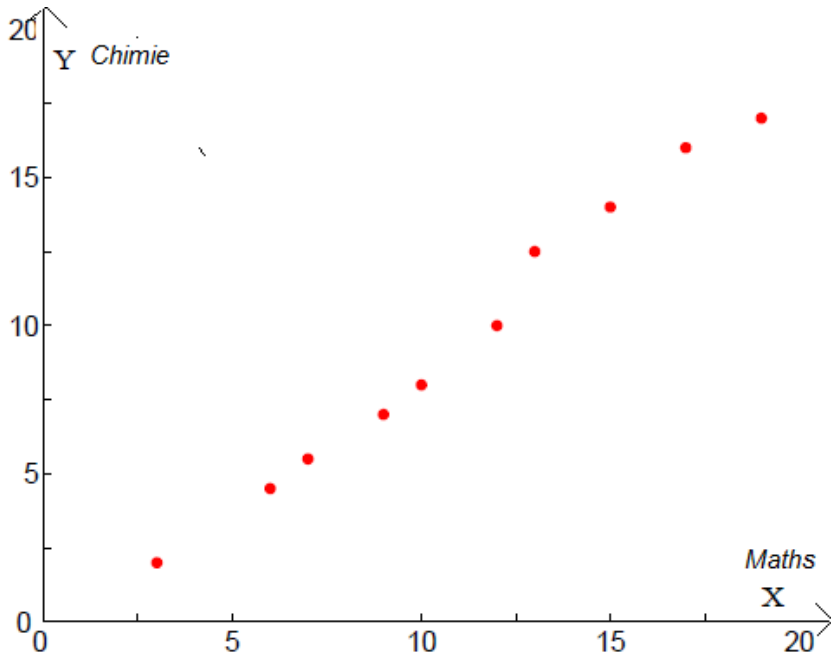


Figure – Le nuage de points $M_i(x_i, y_i)$.

Statistique à deux dimensions

Régression linéaire

On remarque que d'après la répartition des points sur le graphe, une droite est une bonne approximation de ce nuage de points. Il existe plusieurs méthodes pour approximer cette droite, dans ce cours nous nous intéresserons à la méthode des moindres carrées, connue par sa bonne précision.

Statistique à deux dimensions

Méthode de Moindres carrés

La droite de régression a pour equation $y = ax + b$. On doit déterminer les deux paramètres a et b telle que la variation résiduelle des moindres carrées

$$\xi(a;b) = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(y_j - (ax_i + b))^2$$

soit minimal.

On va essayer de démontrer d'où vient la relation de cette approximation (**cette démonstration est hors programme, c'est seulement pour avoir une idée**). $\xi(a;b)$ sera minimum lorsque les dérivées partiels $\frac{\partial \xi}{\partial a} = \frac{\partial \xi}{\partial b} = 0$.

Statistique à deux dimensions

Méthode de Moindres carrés

On calcule

$$\frac{\partial \xi}{\partial a} = -2\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i(y_j - (ax_i + b))$$

$$\frac{\partial \xi}{\partial b} = -2\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(y_j - (ax_i + b))$$

En distribuant l'opérateur $\sum$, il vient

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}y_j x_i - a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i^2 - b\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i = 0$$

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}y_j - a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i - b\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = 0$$

Statistique à deux dimensions

Méthode de Moindres carrés

Ce qui conduit à deux équations dites normales

$$\begin{aligned} a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i^2 + b\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i &= \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}y_j x_i \\ a\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i + b &= \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}y_j \end{aligned}$$

La droite de régression estimée est donc $y = \widehat{a}x + \widehat{b}$ avec $\widehat{a}$ est l'estimateur de a et $\widehat{b}$ est l'estimateur de b qui sont données par :

$$\boxed{\widehat{a} = \frac{Cov(X,Y)}{\sigma_X^2}} \qquad \boxed{\widehat{b} = \overline{Y} - \widehat{a}\overline{X}}$$

Statistique à deux dimensions

Méthode de Moindres carrés

Étant donné un couple (X, Y) de séries quantitatives définies sur une même population Ω, il n'existe qu'une seule manière d'effectuer une régression de Y en X, à l'aide d'une fonction affine f, par la méthode des moindres carrés, la fonction $f(X) = aX + b$ est définie par

$$\widehat{a} = \frac{Cov(X,Y)}{\sigma_X^2}, \qquad \widehat{b} = \overline{Y} - \widehat{a}\overline{X} \text{ et on a } f(\overline{X}) = \overline{Y}$$

Statistique à deux dimensions

Méthode de Moindres carrés

Donc pour tracer une droite d'ajustement par la méthode des moindres carrées, il suffit de tracer la droite

$$Y = aX + b,$$

avec

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

et

$$b = \overline{Y} - \widehat{a}\overline{X}$$

Statistique à deux dimensions

Méthode de Moindres carrés

La droite de régression de la note de Maths en fonction de la note de Chimie, par la méthode des moindres carrés, a pour équation :

$$Y = 0.999X + 1.447,$$

car on a $\overline{X} = 11.1$, $\overline{Y} = 9.65$, $Cov(X, Y) = 23,08$ et $\sigma_X^2 = V(X) = 23,09$. On peut maintenant tracer la courbe de l'ajustement linéaire comme suit :

Statistique à deux dimensions

Méthode de Moindres carrés

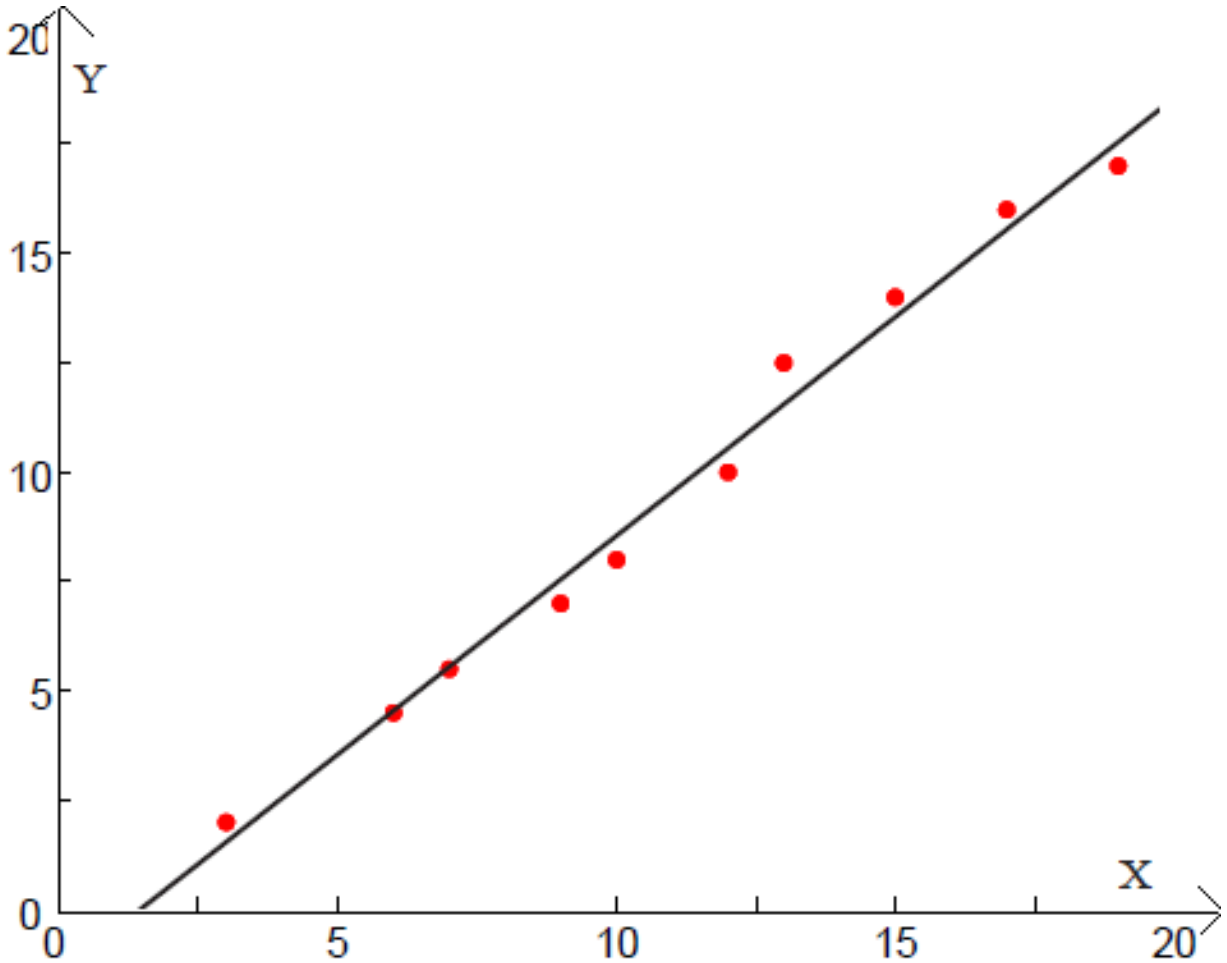


Figure – La courbe d'ajustement linéaire par les moindres carrées.

Statistique à deux dimensions

Méthode de Moindres carrés

Definition

La droite d'équation $y = \frac{Cov(X,Y)}{\sigma_X^2}\left(x - \overline{X}\right) + \overline{Y}$ s'appelle la droite d'ajustement affine de Y en X.

Statistique à deux dimensions

Autres ajustements

Lorsque le nuage est « allongé » mais a une autre forme, on a recours à un ajustement différent ; il en existe de nombreux, reposant sur différentes familles de fonctions : les fonctions logarithmes, exponentielles, puissances, logistiques, etc.

Statistique à deux dimensions

Ajustement exponentiel

our les phénomènes à croissance forte ou à décroissance rapide, il peut être pertinent d'approcher la forme générale du nuage par une fonction exponentielle de la forme :

$$x \longmapsto \exp(ax+b) \text{ ou } x \longmapsto \alpha \exp(x) + \beta$$

Pour déterminer les valeurs de a et b , ou de α et β, on effectue un ajustement affine (par exemple, par la méthode des moindres carrés) sur la série $(x_i,\ \ln(y_i))$ ou la série $(\exp(x_i),\ y_i)$.

Statistique à deux dimensions

Ajustement exponentiel

Le tableau ci-dessous indique le salaire brut annuel, en euros, perçu par un cadre.

année	2005	2006	2007	2008	2009
rang x_i de l'année	1	2	3	4	5
salaire y_i (en €)	42 900	54 200	64 100	81 600	102 000
$z_i = \ln(y_i)$	10,666	10,900	11,068	11,310	11,533

Statistique à deux dimensions

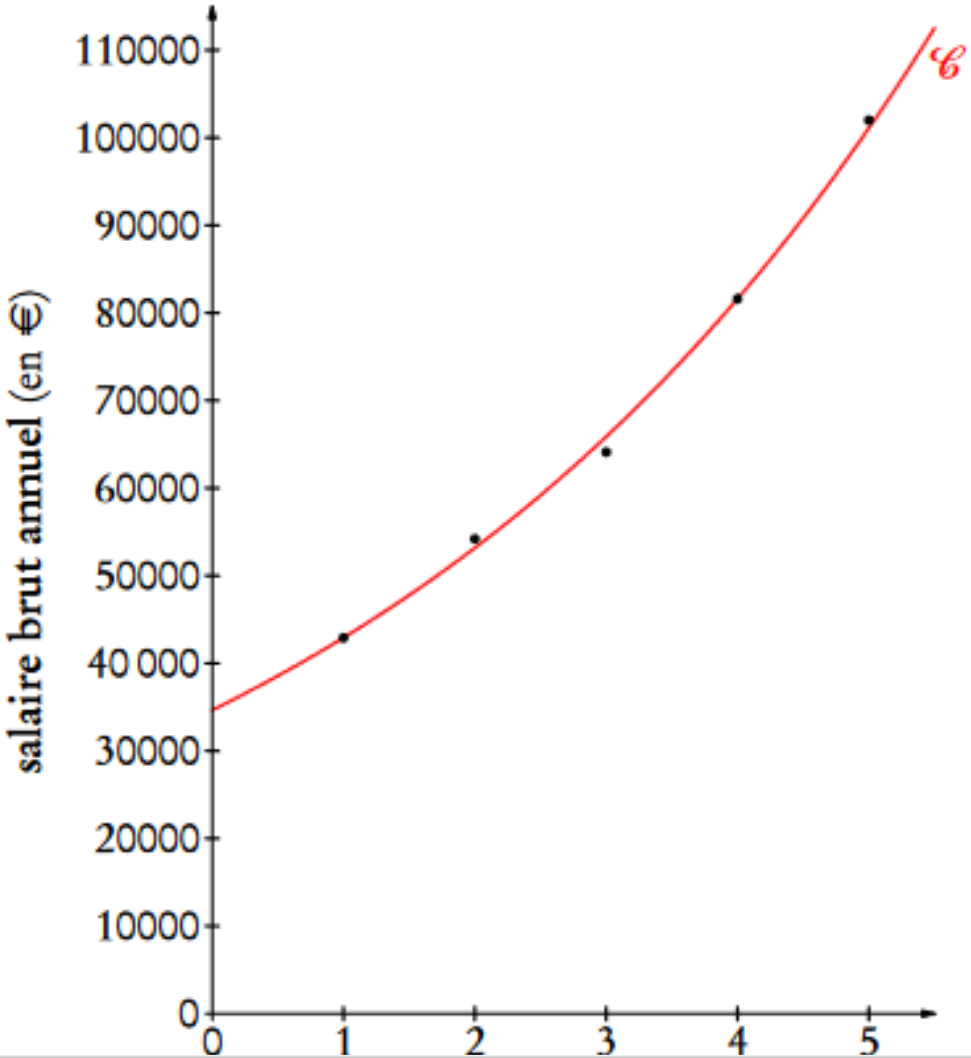
Ajustement exponentiel

Le coefficient de corrélation linéaire entre x et $z = ln(y)$ est très bon ($r \simeq 0,999$), ce qui justifie un ajustement affine. Par la méthode des moindres carrés, on trouve une droite de régression Dx de z en x, d'équation

$$z = 0,21x + 10,45.$$

Statistique à deux dimensions

Ajustement exponentiel



Statistique à deux dimensions

Ajustement exponentiel

$$\ln(y) = 0,21x + 10,45 \Leftrightarrow y = \exp(0,21x + 10,45)$$

ce qui implique :

$$y = 34544 \exp(0,21x).$$

Statistique à deux dimensions

Ajustement logarithmique

À l'opposé des fonctions exponentielles, les fonctions logarithmes :

$$x \longmapsto a\ln(x) + b \text{ ou } x \longmapsto \ln(\alpha x + \beta)$$

sont tout à fait indiquées dans la modélisation des phénomènes à (dé)croissance lente. Pour cela, on effectue un ajustement affine sur la série $(\ln(x_i), y_i)$ ou $(x_i, \exp(y_i))$.

Statistique à deux dimensions

Ajustement logarithmique

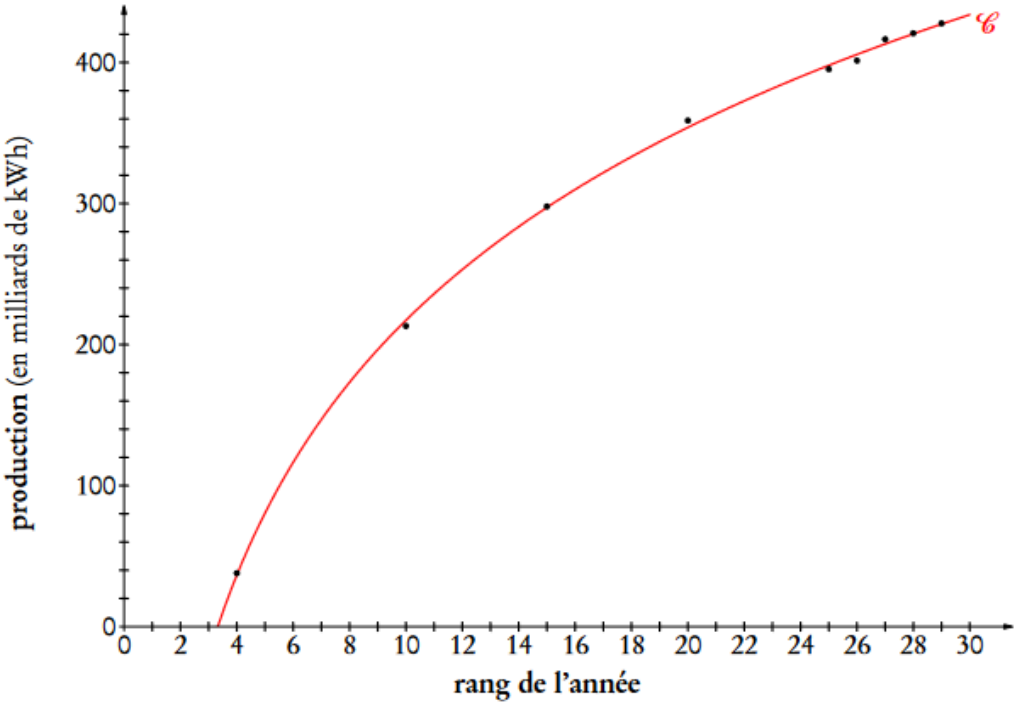
Exemple :

Le tableau ci-dessous donne la production d'électricité d'origine nucléaire en France, exprimée en milliards de kWh, entre 1979 et 2004 (source : site web du Ministère de l'industrie). Les rangs des années sont calculés par rapport à l'année 1975.

année	1979	1985	1990	1995	2000	2001	2002	2003	2004
rang x_i de l'année	4	10	15	20	25	26	27	28	29
production y_i	37,9	213,1	297,9	358,8	395,2	401,3	416,5	420,7	427,7

Statistique à deux dimensions

Ajustement logarithmique



Statistique à deux dimensions

Ajustement logarithmique

La distribution des points du nuage suggère un ajustement logarithmique. Pour l'effectuer, nous allons travailler sur la série $(z_i; y_i)$, où $zi = ln(xi)$.

$z_i = \ln(x_i)$	1,386	2,303	2,708	2,996	3,219	3,258	3,296	3,332	3,367
production y_i	37,9	213,1	297,9	358,8	395,2	401,3	416,5	420,7	427,7

Le coefficient de corrélation linéaire entre z et y est excellent ($r \simeq 0,9997$); on procède donc à un ajustement affine entre z et y. La méthode des moindres carrés fournit une équation de la droite de régression Dz de y en z :

$$y = 197,2z - 236,9$$