Université de Lille

Statistique Descriptive

G. Marot-Briend
guillemette.marot@univ-lille.fr

2021-2022

Présentation des UE

Analyses statistiques univariées et bivariées

- UE1 : commune à MISO et OSB
- UE2 : spécifique à MISO : compléments de cours, logiciel R

Cours communs :

- statistique descriptive
- variables aléatoires, lois usuelles
- synthèse intervalles de confiance et tests usuels
- corrélation régression
- analyse de la variance

Présentation des UE

Planning et supports de cours disponibles sur **Moodle**

Clé d'inscription MISO : 29siw5
Clé d'inscription OSB : di33fs

Evaluation
UE1 : 30% CC + 70% CT
UE2 (MISO) : 50% TP + 50% CT
Remarque : les CT UE1+UE2 ont lieu le même jour (durée totale épreuve CT 2h pour les MISO, 1h30 pour les OSB)

Introduction

STATISTIQUE :

Méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux puis à analyser, commenter, critiquer ces données

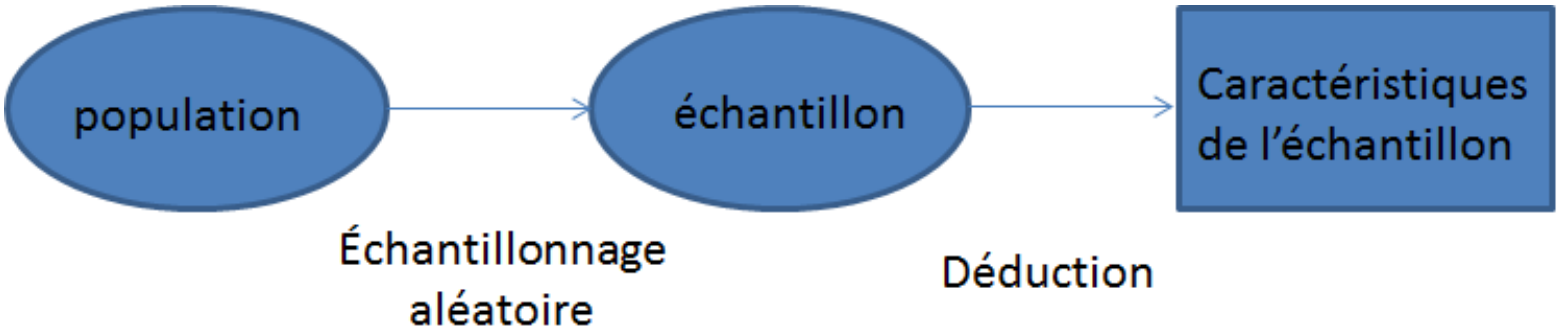
Ne pas confondre la statistique (science) avec une statistique (ensemble de données chiffrées sur un sujet précis)

Introduction

Etapes d'une étude statistique

- collecte des données issues de l'observation ou de l'expérimentation
- analyse statistique
  - analyse descriptive : résumer et présenter les données observées
  - inférence : étendre ou généraliser les conclusions obtenues

Statistique descriptive



- population : ensemble étudié ; les éléments de la population sont appelés individus ou unités statistiques

- échantillonnage : ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

L'échantillonnage aléatoire simple consiste à prélever au hasard et de façon indépendante n individus d'une population à N individus.

Plan

Types de variables

Variables qualitatives

- nominales : variables à plusieurs modalités non mesurables qui s'excluent mutuellement.
  Ces modalités sont exprimables par des noms et ne sont pas hiérarchisées.

  Les individus sont caractérisés par leur appartenance à une seule des modalités que peut prendre la variable.
  ex : couleur de cheveux $\Rightarrow$ blond, brun, noir,...

  Dans le cas de deux modalités, on parle de variables binaires ou dichotomiques.
  ex : fumeur (oui/non)

Types de variables

Variables qualitatives

- ordinales :
  Les modalités traduisent le degré d'un état caractérisant un individu sans que ce degré ne puisse être défini par un nombre qui résulte d'une mesure.
  Les modalités sont hiérarchisées.
  ex : le stade d'une maladie

Types de variables

Variables quantitatives

Les variables quantitatives sont le résultat d'une mesure ou d'un comptage. On distingue deux catégories secondaires :

- les variables discrètes : elles ne peuvent prendre qu'un nombre fini de valeurs, elles sont souvent issues d'un comptage.
  ex : le nombre d'enfants dans une famille
- les variables continues : elles peuvent prendre un nombre infini de valeurs, elles sont souvent issues d'une mesure effectuée avec un instrument spécifique dans une unité et avec une précision adaptée.
  ex : âge, taille

Types de variables

Remarques :

En réalité, le nombre de valeurs possibles pour un caractère donné dépend de la précision de la mesure.

On peut considérer comme continu un caractère discret qui peut prendre un grand nombre de valeurs

ex : nombre de globules blancs ou rouges par mL de sang.

Plan

Représentation des données

Une série statistique correspond aux différentes modalités d'un caractère sur un échantillon d'individus appartenant à une population donnée.

Ex : groupe sanguin $\Rightarrow$ A AB O O AB A B A AB AB A ...

Représentation des données

Tableaux de distribution

- variables qualitatives nominales

Pour chaque modalité, on détermine l'effectif n_i, c'est à dire le nombre de sujets présentant la modalité.
Les modalités doivent être mutuellement exclusives $\Rightarrow$ l'effectif total du groupe étudié est égal à la somme des effectifs de chaque modalité :

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

avec p le nombre de modalités de la variable et n l'effectif total.

On rassemble les résultats d'une distribution de fréquences sous forme d'un tableau.

Tableaux de distribution

Exemple : tableau de répartition des groupes sanguins dans un hôpital du Nord Pas de Calais

groupe	effectif (n_i)	fréquence (f_i)
O	45	0,45
A	40	0,40
B	10	0,10
AB	5	0,05
Total	100	1

On appelle fréquence de la modalité x_i $f_i = n_i/n$.
Un pourcentage est une fréquence exprimée en %, cad $100 f_i$.

Tableaux de distribution

Les variables qualitatives ne permettent pas le calcul des paramètres statistiques usuels (moyenne, variance, ...)

Deux variables qualitatives P et T peuvent être mesurées sur le même individu. Les valeurs obtenues sont placées dans un tableau à double entrée dit tableau de contingence.

Tableaux de distribution

Tableau de contingence

n_{ij} : effectif des individus possédant à la fois la modalité de la ligne i et de la colonne j

		T			somme
P	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1t}	$n_{1.}$
	.		.		.	.
	.		.		.	.
	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{it}	$n_{i.}$
	.		.		.	.
	.		.		.	.
	n_{p1}	$\dots$	n_{pj}	$\dots$	n_{pt}	$n_{p.}$
somme	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.t}$	$n_{..}$

Représentation des données

Tableaux de distribution

- variables quantitatives discrètes et variables qualitatives ordinales

On appelle fréquences cumulées croissantes

$$F_i = \sum_{k=1}^{i} f_k$$

et fréquences cumulées décroissantes

$$G_i = \sum_{k=i}^{p} f_k$$

Ceci a un sens pour les variables quantitatives discrètes et ordinales puisqu'on peut ordonner les modalités. On appelle $x_1, \ldots, x_i, \ldots, x_p$ les p valeurs ordonnées de x (l'indice i correspond alors au rang).

Représentation des données

Tableau de distribution des effectifs et fréquences cumulés

valeur du caractère modalité x_i	effectif n_i	fréquence f_i	effectif cumulé N_i	fréquence cumulée F_i
x_1	n_1	f_1	n_1	f_1
x_2	n_1	f_2	$n_1 + n_2$	$f_1 + f_2$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_i	n_i	f_i	$n_1 + n_2 + \ldots + n_i$	$f_1 + f_2 + \ldots + f_i$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_p	n_p	f_p	n	1

Ex : nombre d'enfants dans les familles

Représentation des données

Nombre d'enfants dans les familles :

xi	ni	fi	Ni	Fi	Gi
0	10	0,1			
1	24	0,24			
2	32				
3	19		85	0,85	0,34
4	8		93	0,93	0,15
5	4				
6			100		

Représentation des données

Nombre d'enfants dans les familles :

xi	ni	fi	Ni	Fi	Gi
0	10	0,1	10	0,1	1
1	24	0,24	34	0,34	0,9
2	32	0,32	66	0,66	0,66
3	19	0,19	85	0,85	0,34
4	8	0,08	93	0,93	0,15
5	4	0,04	97	0,97	0,07
6	3	0,03	100	1	0,03

Représentation des données

Tableaux de distribution

- variables quantitatives continues

Il est nécessaire de regrouper en classes les valeurs prises par la variable.

Ex : taille (en cm) [150-160[,[160-170[, [170-180[

L'intervalle de classe, également appelé amplitude, est la différence entre la borne supérieure et la borne inférieure.

En règle générale, on choisit des classes de même amplitude.

Représentation des données

Si l'amplitude n'est pas constante, il faut calculer la densité de fréquence :

$$d_i = \frac{f_i}{\text{amplitude}_i}$$

La densité de fréquence permet de comparer les fréquences d'une classe à l'autre.

Exercice : taille

Représentation des données

Tailles des individus en cm :

Classe	Ci	ni	fi	di	Ni	Fi
$[140 - 160[$		10				
$[160 - 165[$		20				
$[165 - 170[$		30				
$[170 - 175[$	172.5	45	0,225	0,045	105	0,525
$[175 - 180[$		40			145	0,725
$[180 - 185[$		35			180	0,9
$[185 - 190[$		15	0,075	0,015		
$[190 - 200[$		5	0,025	0,0025		

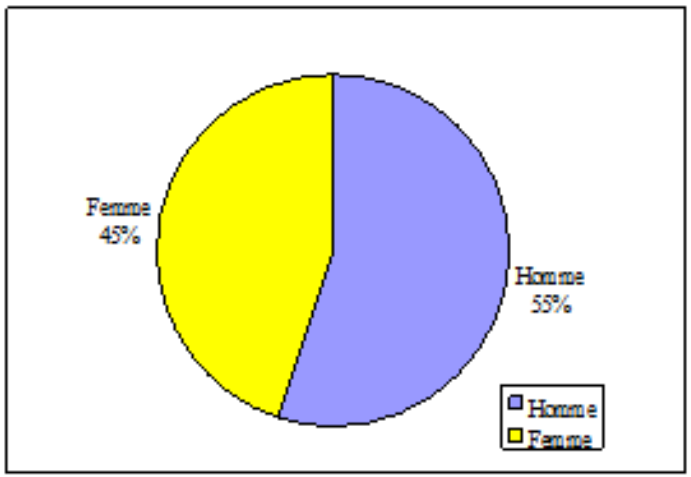
Représentation des données

Tailles des individus en cm :

Classe	Ci	ni	fi	di	Ni	Fi
$[140 - 160[$	150	10	0,05	0,0025	10	0,05
$[160 - 165[$	162,5	20	0,1	0,02	30	0,15
$[165 - 170[$	167,5	30	0,15	0,03	60	0,3
$[170 - 175[$	172,5	45	0,225	0,045	105	0,525
$[175 - 180[$	177,5	40	0,2	0,04	145	0,725
$[180 - 185[$	182,5	35	0,175	0,035	180	0,9
$[185 - 190[$	187,5	15	0,075	0,015	195	0,975
$[190 - 200[$	195	5	0,025	0,0025	200	1

Représentations graphiques

Variables qualitatives



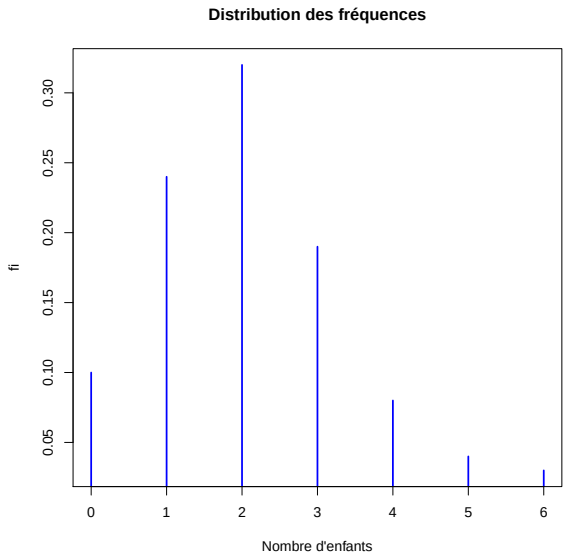
VARIABLES QUALITATIVES

Angle α_i du $i^{\text{ème}}$ groupe :

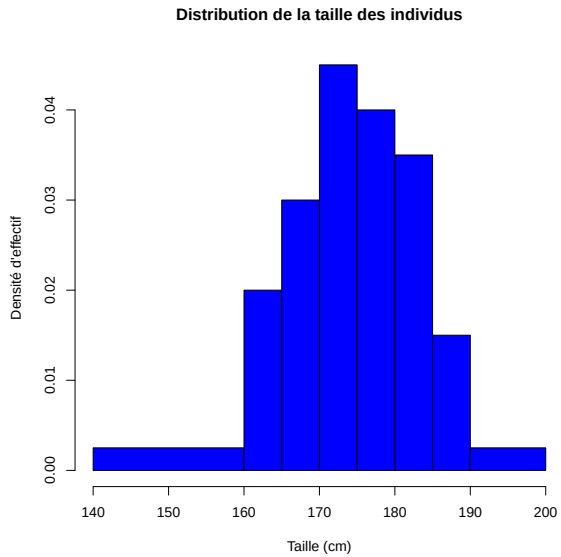
$$\alpha_i = 360 * \frac{n_i}{n} = 360 f_i$$

Représentations graphiques

Variables quantitatives



Var. Discretes



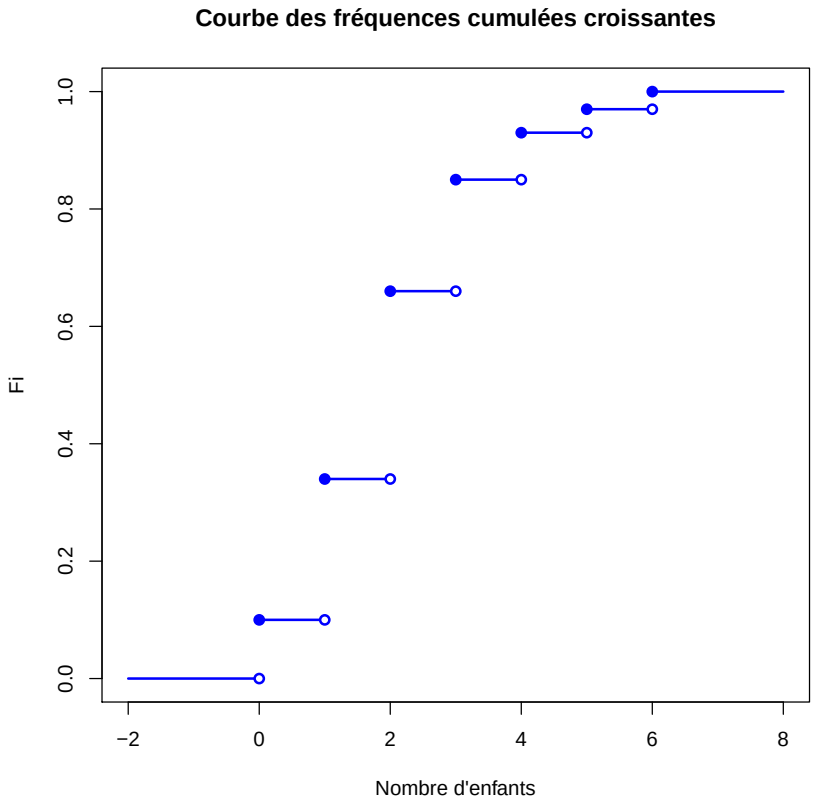
Var continues

Pour l'histogramme, l'aire dans chaque rectangle doit être proportionnelle à l'effectif de la classe.
Si les amplitudes sont égales, la hauteur du rectangle peut être égale à l'effectif, sinon, elle doit être égale à la densité d'effectif (ou de fréquence).

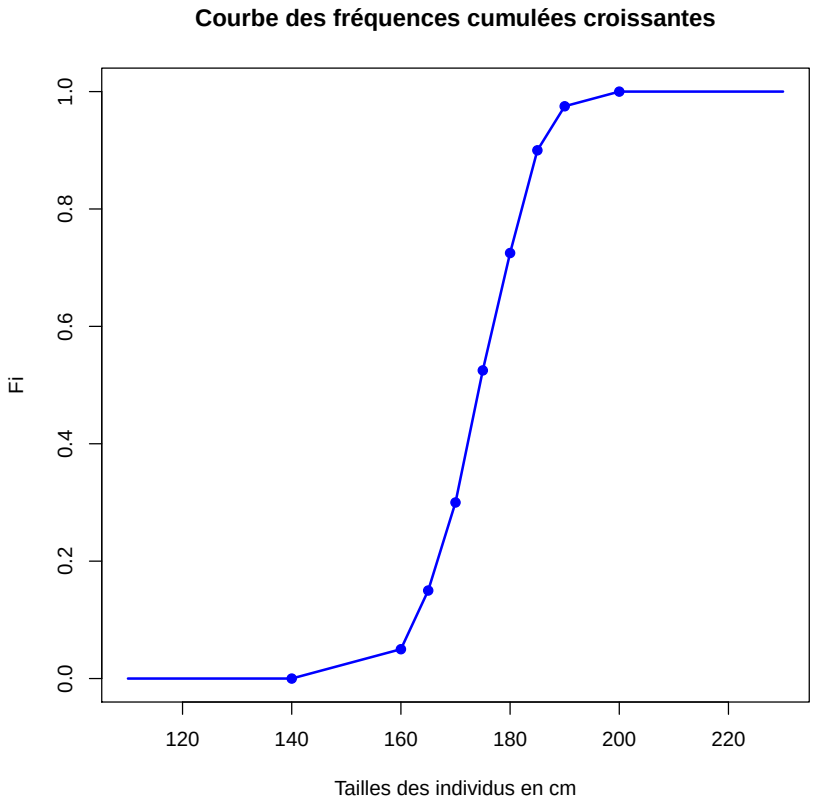
Représentations graphiques

Courbes des fréquences cumulées croissantes

Variable discrète



Variable continue



Plan

Indicateurs numériques

Indicateurs numériques (seulement pour les variables quantitatives)

- indicateurs de position : mode, moyenne, médiane, quartiles

Le mode d'une distribution est la valeur la plus fréquente de celle-ci.

Exercice : Trouver le mode de la distribution du nombre d'enfants

Indicateurs numériques

Indicateurs numériques (seulement pour les variables quantitatives)

- indicateurs de position : mode, moyenne, médiane, quartiles

Le mode d'une distribution est la valeur la plus fréquente de celle-ci.

Exercice : Trouver le mode de la distribution du nombre d'enfants

Si les données sont regroupées par classe, on définit la classe modale comme la classe dont la densité d'effectif est la plus élevée et on attribue (arbitrairement) au mode la valeur centrale de cette classe.

Exercice : Tracer des exemples de densités de distributions unimodale et bimodale.

Indicateurs de position

Moyenne arithmétique

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

- variable discrète

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

avec n_i l'effectif correspondant à la valeur x_i

- variable continue

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i c_i}{n}$$

avec n_i l'effectif de la classe i et c_i la valeur centrale de la classe i.

Indicateurs de position

Propriétés

- $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- si $y = ax + b$ alors $\bar{y} = a\bar{x} + b$

Remarques :

- la moyenne est très sensible aux valeurs extrêmes.
- la moyenne d'un groupe résultant de la fusion d'autres groupes n'est égale à la moyenne des moyennes que si tous les groupes ont le même effectif.
- il existe d'autres moyennes (géométrique, harmonique, quadratique) peu utilisées en santé.

Indicateurs de position

Médiane M_e est la valeur qui partage la série classée en deux groupes de même effectifs. C'est aussi la valeur du caractère pour laquelle la fréquence cumulée est égale à 0.5.

Exercice : Trouver la médiane de la série $\{3,6,4,10,8\}$

Indicateurs de position

Médiane M_e est la valeur qui partage la série classée en deux groupes de même effectifs. C'est aussi la valeur du caractère pour laquelle la fréquence cumulée est égale à 0.5.

Exercice : Trouver la médiane de la série $\{3,6,4,10,8\}$

En pratique

- série impaire constituée de $2k+1$ éléments
  $\Rightarrow M_e$ $k+1^{\text{ème}}$ valeur
- série paire constituée de $2k$ éléments

$$M_e = \frac{x_k + x_{k+1}}{2}$$

Indicateurs de position

Médiane M_e est la valeur qui partage la série classée en deux groupes de même effectifs. C'est aussi la valeur du caractère pour laquelle la fréquence cumulée est égale à 0.5.

Exercice : Trouver la médiane de la série $\{3,6,4,10,8\}$

En pratique

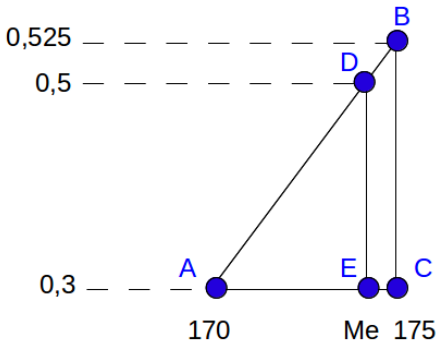
- série impaire constituée de $2k+1$ éléments
  $\Rightarrow M_e$ $k+1^{\text{ème}}$ valeur
- série paire constituée de $2k$ éléments

$$M_e = \frac{x_k + x_{k+1}}{2}$$

Quand la taille de l'échantillon est important, on utilise soit un graphique soit le calcul par interpolation linéaire.

Indicateurs de position

Exemple : Taille



Théorème de Thalès :

$$\frac{AD}{AB} = \frac{AE}{AC} = \frac{DE}{BC}$$

$$\frac{AE}{AC} = \frac{DE}{BC}$$

$$\frac{Me - x_{\text{inf i}}}{l_i} = \frac{0,5 - F_{\text{inf i}}}{f_i}$$

$$\frac{Me - 170}{175 - 170} = \frac{0,5 - 0,3}{0,525 - 0,3}$$

$$Me = 170 + (175 - 170)\frac{0,5 - 0,3}{0,525 - 0,3}$$

$$Me = 174,4$$

$$M_e = x_{\text{inf i}} + l_i \left(\frac{0,5 - F_{\text{inf i}}}{f_i} \right)$$

Indicateurs de position

Les quartiles sont les valeurs qui partagent la série ordonnée en 4 groupes de même effectif.

Les percentiles sont les valeurs qui partagent la série ordonnée en 100 groupes de même effectif.

Remarques :

- Le deuxième quartile correspond à la médiane
- 50% des individus ont des mesures comprises entre le 1^{er} et le $3^{ème}$ quartile.
- Les percentiles sont utilisés en médecine, notamment dans les courbes de croissance des nouveaux nés.
- Le 1^{er} percentile est la valeur x_i telle que 1% des individus aient des valeurs inférieures.

Indicateurs de position

Comparaison des caractéristiques du mode, de la moyenne et de la médiane

	Avantages	Inconvénients
Mode	. . .	. . .
Moyenne	. . .	. . .
Médiane	. . .	. . .

Indicateurs de position

MODE

Avantages :

- caractéristique intéressante dans le cas des distributions asymétriques
- bon indicateur de population hétérogène
- non influencé par les valeurs extrêmes

Inconvénients :

- se prête mal aux calculs statistiques
- très sensible aux variations d'amplitude des classes
- son calcul ne tient compte que des individus dont les valeurs se rapprochent de la classe modale

Indicateurs de position

MOYENNE

Avantages :

- se prête facilement aux calculs et aux tests statistiques
- elle a d'autant plus de sens que la répartition est symétrique et la dispersion plus faible

Inconvénients :

- très sensible aux valeurs extrêmes
- ne convient pas aux valeurs qualitatives ordinales
- représente mal une population hétérogène (polymodale)

Indicateurs de position

MEDIANE

Avantages :

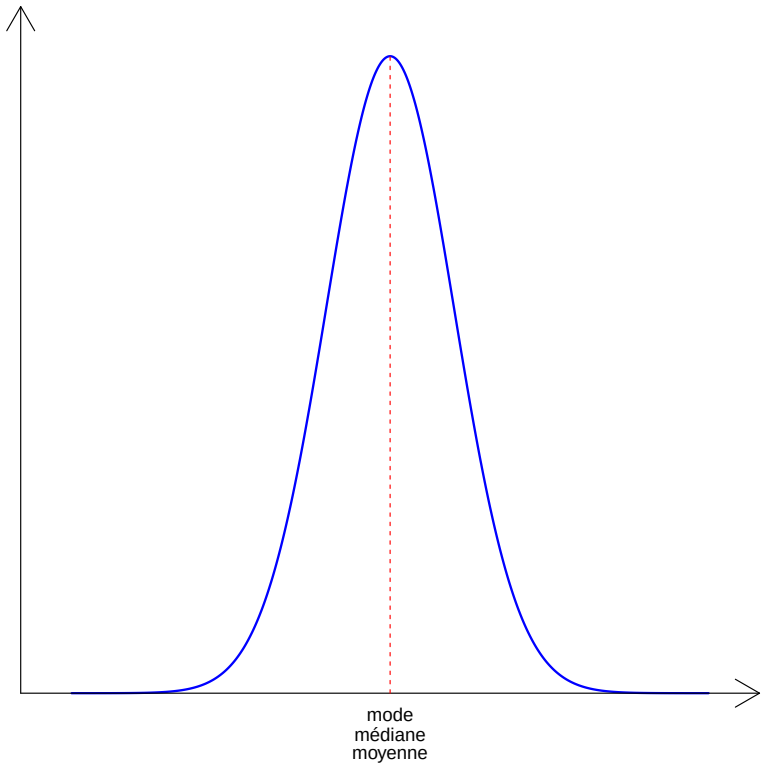
- moins sensible que la moyenne aux valeurs extrêmes
- peut être utilisée avec des variables ordinales
- peu sensible aux variations d'amplitude des classes

Inconvénients :

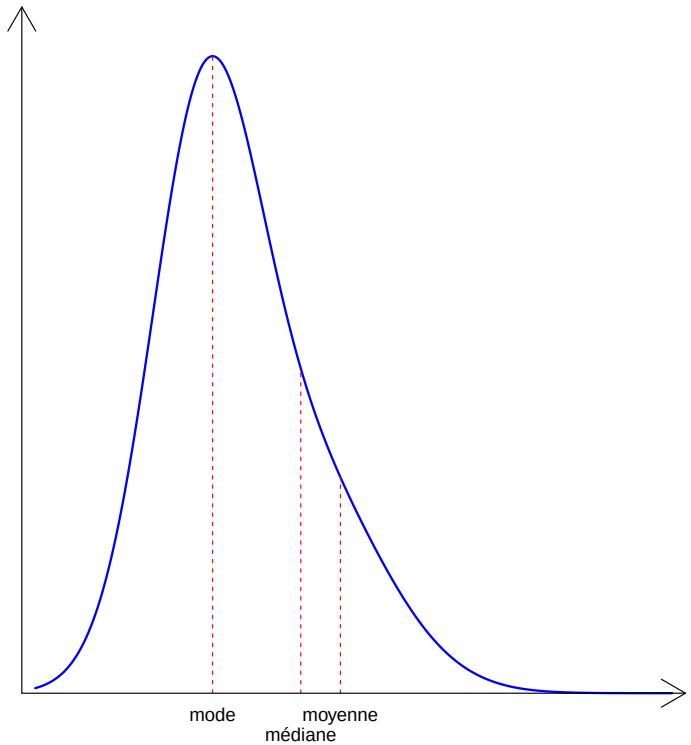
- se prête mal aux calculs statistiques
- suppose l'équirépartition des données
- le classement peut être long si les valeurs sont nombreuses

Indicateurs de position

Distribution symétrique



Distribution asymétrique



Indicateurs numériques

Indicateurs numériques (seulement pour les variables quantitatives)

- indicateurs de dispersion :

Paramètres peu utilisés :

- valeurs extrêmes : plus petite et plus grande des valeurs
- étendue : différence entre les valeurs extrêmes
- écart absolu moyen : somme des écarts à la moyenne en valeur absolue divisée par l'effectif total $1/n \sum |x_i - \bar{x}|$
- distance inter-quartiles différence entre les $3^{\text{ème}}$ et 1^{er} quartiles

Indicateurs de dispersion

Indicateurs de dispersion les plus utilisés

Variance observée

Soit un échantillon de n valeurs observées $x_1, x_2, \ldots, x_n$ d'une variable quantitative X et soit $\bar{x}$ sa moyenne observée.

$$s^2_{\text{ech}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Indicateurs de dispersion

Théorème de König-Huygens :

$$s^2_{\text{ech}} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$$

Dans le cas de données regroupées en k classes d'effectifs n_i,

$$s^2_{\text{ech}} = \frac{1}{n}\sum_{i=1}^{k} n_i (c_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \bar{x}^2$$

Exercice : Calculs de la moyenne et de la variance du nombre d'enfants dans les familles et de la taille des individus.

Exercice

Nombre d'enfants par famille

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{p} n_i x_i}{n} \\
&= \frac{1}{100}[10 * 0 + 24 * 1 + 32 * 2 + 19 * 3 + 8 * 4 + 4 * 5 + 3 * 6] \\
&= \frac{215}{100} \\
\bar{x} &\approx 2
\end{aligned}$$

$$\begin{aligned}
s^2_{\text{ech}} &= \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2 \\
&= \frac{1}{100}[10 * 0^2 + 24 * 1^2 + 32 * 2^2 + \cdots + 3 * 6^2] - \left(\frac{215}{100}\right)^2 \\
&\approx 2
\end{aligned}$$

Exercice

Taille des individus

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{k} n_i c_i}{n} \\ &= \frac{1}{200}[10 * 150 + 20 * 162,5 + 30 * 167,5 + \cdots + 5 * 195] \\ &= \frac{34812,5}{200} \\ \bar{x} &\approx 174,1\text{cm}\end{aligned}$$

$$\begin{aligned}s^2_{\text{ech}} &= \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \bar{x}^2 \\ &= \frac{1}{200}[10 * 150^2 + 20 * 162,5^2 + \cdots + 5 * 195^2] - \left(\frac{34812,5}{200}\right)^2 \\ &\approx 88,7 \text{ unités}\end{aligned}$$

Indicateurs de dispersion

Remarque :

La dimension de la variance est le carré de celle de la variable
$\Rightarrow$ il est difficile d'utiliser la variance comme norme de dispersion car le recours au carré conduit à un changement d'unités.
Elle n'a donc pas de sens biologique direct contrairement à l'écart-type qui s'exprime dans les mêmes unités que la moyenne.

Ecart-type

$$s_{ech} = \sqrt{s_{ech}^2}$$

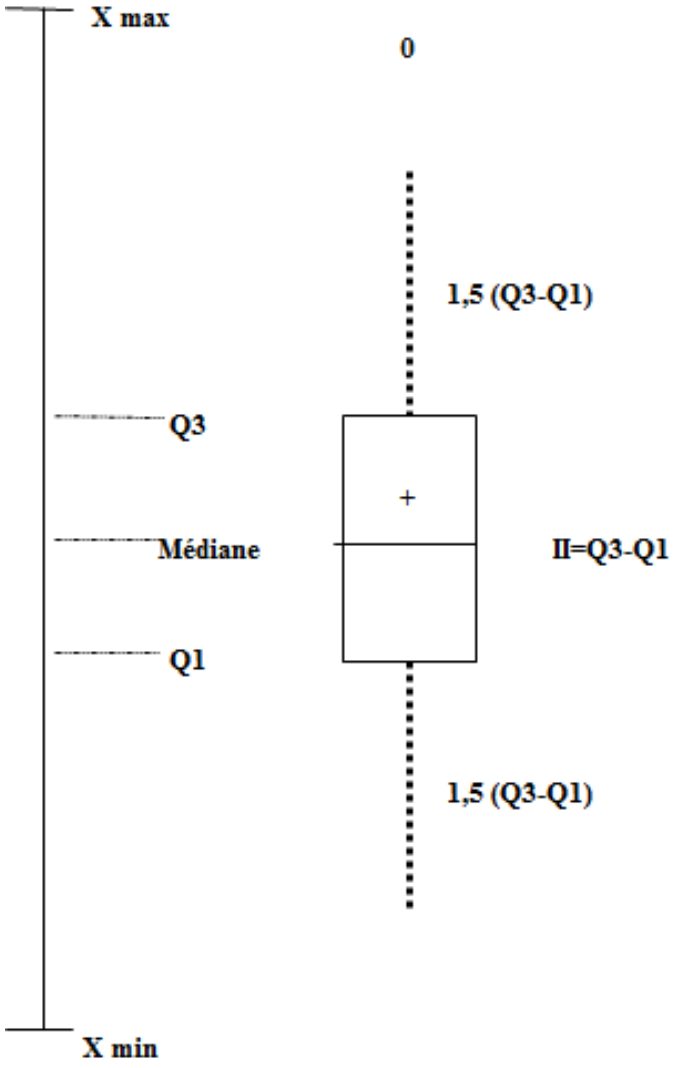
Indicateurs de dispersion

La variance et l'écart-type observés sont des paramètres de dispersion absolue qui mesurent la variation absolue des données indépendamment de l'ordre de grandeur des données.

Le coefficient de variation, noté CV, est un indice de dispersion relatif prenant en compte ce biais :

$$CV = \frac{s_{\text{ech}}}{\bar{x}}$$

Boite à moustaches



Eléments d'une boîte à moustaches

- la "boîte" : rectangle dont la longueur est comprise entre le 1^{er} et le $3^{ème}$ quartiles
  $\Rightarrow$ 50% des valeurs sont dans l'intervalle.
- les "moustaches" : traits verticaux de longueur $1,5 * (Q3 - Q1)$, raccourcis aux minimum et maximum des observations si il n'y a pas de valeurs en dehors des moustaches.
- une ligne à l'intérieur du rectangle : la médiane.