Chapitre 1: Série statistique à une seule variable

Said, El Melhaoui

Faculté des Sciences Juridiques, économiques et Sociales Oujda

http://said-el-melhaoui.e-monsite.com

Outline

Population, unité et caractère

- **population statistique** : Toute ensemble au sens mathématique que l'on soumet à une étude statistique.
- **Unité statistique** (**individu**) : chaque élément de cette population.
- **Caractère** : Un aspect, une propriété, un trait qui est commun à tous les individus et qui est objet de l'étude statistique.

Exemple 1: On veut étudier '*le nombre de ventes journaliers d'un produit dans un magasin pendant le mois de septembre*'.

- Population : c'est l'ensemble des 30 jours
- l'unité statistique: c'est le jour
- le caractère d'intérêt: est le nombre de produits vendus par jour

Population, unité et caractère (suite)

Exemple 2: On veut étudier '*les tailles des sportifs participant à une compétition*'.

- La population: l'ensemble des sportifs participant à la compétition
- l'unité statistique: un sportif
- caractère étudié: c'est la taille du sportif

Exemple 3: Pour étudier '*l'image de marque d'un produit dans le marché, on demande à cent clients d'exprimer leur satisfaction du produit, via l'une des appréciations suivantes: très satisfait, satisfait, pas mal, non satisfait*'.

- population: est l'ensemble des 100 clients
- unité statistique: chaque client
- caractère: l'appréciation du client

Population, unité et caractère (suite)

Exemple 4: On veut étudier '*le groupe sanguin de* 40 *personnes atteintes par l'hémophilie*'.

- population: l'ensemble des 40 malades
- chaque malade
- caractère: est le groupe sanguin

Type de caractère

On distingue plusieurs types de caractère:

- **le Caractère quantitatif**: il prend des valeurs numériques; il est mesurable : le poids, la taille, l'âge, le revenu, la note etc. Il est aussi appelé **variable statistique**.
  Il existe deux sortes de variable statistique:
  - dans le cas ou la variable statistique ne peut prendre qu'un nombre fini de valeurs isolées, alors la variable est dite **variable discrète** (Voir Exemple 1).
  - lorsque les valeurs de la variable statistique peuvent prendre tout réel d'un intervalle de l'ensemble $\mathbb{R}$, la variable est dite **variable continue** (Voir Exemple 2)

Type de caractère (suite)

- **Caractère qualitatif** : il ne peut être exprimé par des nombres, il est non numériquement mesurable. on ne parle pas de valeurs mais de **modalités**.
  On distingue deux types de caractère qualitatif:
  - dans le cas ou les modalités du caractère peuvent être ordonnées suivant un ordre bien défini alors le caractère est dit **ordinal** (voir Exemple 3)
  - dans le cas contraire, c'est à dire lorsque les modalités ne peuvent être ordonnées selon un ordre fiable, le caractère est dit **nominal** (voir Exemple 4).

Méthode d'observation

- L'étude d'un caractère auprès de tous les individus d'une population fini s'appelle **recensement** comme, par exemple, *les Recensements du Maroc 2004, 2014*.
- L'étude complète d'une population n'est pas toujours possible, vu les contraintes de temps, de coût et de la non accessibilité de certaines unités etc.
- On se limite, donc, à l'étude d'une partie de la population dite **échantillon**. On dit qu'on fait **échantillonnage** ou **sondage**.
- Dépouillement des observations: c'est la façon de résumer et de classer les données. Une technique assez répondue consiste en l'utilisation d'un **tableau de pointage**.

Méthode d'observation (suite)

Exemple: On s'interesse au groupe sanguin de 40 malades, les modalités retenues sont A, B, AB et O. Les résultats sont

A	*B*	*B*	*O*	*O*	*O*	*AB*	*A*	*O*	*O*
A	*O*	*O*	*B*	*AB*	*AB*	*B*	*B*	*O*	*A*
A	*O*	*B*	*O*	*O*	*O*	*A*	*AB*	*O*	*B*
O	*A*	*O*	*B*	*AB*	*B*	*AB*	*B*	*A*	*O*

Méthode d'observation (suite)

Le tri à plat des résultats donne lieu au tableau suivant:

MODALITE	*POINTAGE*	EFFECTIF
A	𝍸	8
B	𝍸	10
AB	𝍸 |	6
O	𝍸 𝍸 𝍸 |	16

Série statistique à une seule variable

- Lorsque l'étude statistique d'une population concerne un seul caractère (variable) on parle d'une **Série statistique à une seule variable** (**série statistique univariée**)
- La série univariée est organisée (résumée) dans un tableau dit **tableau des effectifs**.

Variable discrète: effectifs

Définition

Soit **E** une population statistique et x une variable statistique qui prend les valeurs isolées et ordonnées

$$x_1 < x_2 < \ldots < x_p$$

- p est le nombre des valeurs prises pa x ;
- n_i **effectif** de x_i est le nombre d'individus où x prend la valeur x_i ;
- L'ensemble des couples $(x_i, n_i)_{1 \leq i \leq p}$ est appelée **série statistique univariée** ;
- **L'effectif total** n est le nombre des individus constituant **E** :

$$n := \sum_{i=1}^{p} n_i = n_1 + n_2 + \ldots + n_p.$$

Variable discrète : Exemple

Exemple 5:
Lors d'une enquête on a interrogé 50 employés afin de connaître le nombre de personnes qu'ils avaient à charge.
Les **données brutes** sont:

0	3	1	4	3	0	4	1	3	1	5	2	4	2	3
3	2	5	5	2	4	2	2	2	4	1	1	2	3	5
1	0	3	3	4	5	1	2	1	2	3	2	2	2	4
0	3	0	2	2										

Les données ordonnées sont

0	0	0	0	0	1	1	1	1	1	1	1	1	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	3	3
3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
5	5	5	5	5										

Variable discrète : Exemple (suite)

L'effectif total est $n = 50$
La série ordonnée est

$$x_1 = 0, x_2 = 1, x_3 = 2x_4 = 3, x_5 = 4, x_6 = 5.$$

Le nombre des valeurs prises par x est $p = 6$
Le tableau des effectifs est :

Valeurs x_i	Effectifs n_i
0	5
1	8
2	15
3	10
4	7
5	5
Total	50

Variable discrète: fréquences

Definition

Soit $(x_i, n_i)_{1\leq i\leq p}$ une série statistique, associée à une population **E** d'effectif total n.

- **La fréquence relative** f_i de la valeur x_i est la proportion des individus ayant le caractère x_i dans la population **E**:

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

- **Le pourcentage** p_i de la valeur x_i est la fréquence multipliée par 100:

$$p_i := f_i \times 100.$$

Variable discrète: Exemple

Les fréquences et les pourcentages de la série des '*personnes à charge*' sont :

Valeurs x_i	Fréquences f_i	Pourcentages p_i
0	0.10	10%
1	0.16	16%
2	0.30	30%
3	0.20	20%
4	0.14	14%
5	0.10	10%
Total	1	100%

Remarque

La fréquence relative et le pourcentage jouent le même rôle, sauf que la dernière est agrandie en échelle pour éliminer la virgule et faciliter l'interprétation.

Variable discrète: cumules

Definition

Soit $(x_i, n_i)_{1\leq i\leq p}$ une série statistique, associée à une population **E** d'effectif total n.

- **L'effectif cumulé** N_i associé à la valeur x_i est l'effectif de toute les valeurs de la variable inférieures ou égales à x_i :

$$N_i := \sum_{j=1}^{i} n_j = n_1 + n_2 + \ldots + n_i.$$

- **La fréquence cumulée** F_i de la valeur x_i est la fréquence de toute les valeurs du caractère inférieures ou égales à x_i :

$$F_i := \sum_{j=1}^{i} f_j = f_1 + f_2 + \ldots + f_i = \frac{N_i}{n}.$$

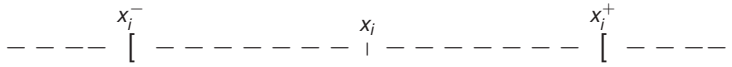
Variable discrète: Exemple

Les effectifs cumulés et les fréquences cumulées de la série des '*personnes à charge*' sont :

Valeurs x_i	Effectifs n_i	Fréquences f_i	Effectifs cumulés N_i	Fréquences cumulées F_i
0	5	0.10	5	0.10
1	8	0.16	13	0.26
2	15	0.30	28	0.56
3	10	0.20	38	0.76
4	7	0.14	45	0.90
5	5	0.10	50	1
Total	50	1	–	–

Variable continue: classes

Lorsque la variable x est continue, on regroupe les valeurs de cette variable dans des intervalles I_i, $i = 1, \ldots, p$ dits **classes statistiques** de la forme :



- p : le nombre de classes retenues ;
- $I_i := [x_i^-, x_i^+[$: la ième classe ;
- $a_i := x_i^+ - x_i^-$: **l'amplitude** de I_i ;
- $x_i := (x_i^+ - x_i^-)/2$: le **centre** de I_i ;
- n_i : **effectif** de la classe I_i, c'est le nombre d'individus ayant une valeur de la variable incluse dans l'intervalle $[x_i^-, x_i^+[$.

Variable continue: classes (suite)

Comment construire les classes?

- Logiquement, le minimum est inclus dans la première classe et le maximum est inclus dans la dernière ;
- de préférence on utilise des classes de même amplitudes
- Le choix du nombre de classes p peut se faire selon la **règle de Sturges** :

$$p \approx 1 + \log_2(n).$$

Variable continue: Exemple

Exemple 2 (suite) : Les tailles exprimées en mètres des 120 sportifs sont :

1.85	1.60	1.85	1.60	1.86	1.86	1.70	1.78	1.79	1.71	1.71	1.79
1.73	1.73	1.61	1.74	1.74	1.88	1.75	1.75	1.75	1.64	1.76	1.76
1.85	1.85	1.73	1.61	1.72	1.62	1.81	1.81	1.73	1.82	1.82	1.64
1.83	1.84	1.83	1.83	1.84	1.84	1.68	1.77	1.69	1.69	1.69	1.70
1.89	1.89	1.90	1.81	1.66	1.78	1.78	1.70	1.78	1.79	1.79	1.79
1.72	1.80	1.72	1.90	1.81	1.81	1.62	1.67	1.64	1.75	1.64	1.82
1.80	1.72	1.81	1.72	1.61	1.77	1.72	1.72	1.81	1.73	1.76	1.76
1.65	1.65	1.65	1.66	1.91	1.66	1.67	1.63	1.67	1.67	1.68	1.68
1.76	1.87	1.87	1.77	1.88	1.75	1.88	1.71	1.79	1.71	1.79	1.79
1.76	1.86	1.77	1.88	1.77	1.72	1.71	1.69	1.77	1.78	1.77	1.71

Variable continue: Exemple (suite)

La série ordonnée

1.60	1.60	1.61	1.61	1.61	1.62	1.62	1.63	1.64	1.64	1.64	1.64
1.65	1.65	1.65	1.66	1.66	1.66	1.67	1.67	1.67	1.67	1.68	1.68
1.68	1.69	1.69	1.69	1.69	1.70	1.70	1.70	1.71	1.71	1.71	1.71
1.71	1.71	1.72	1.72	1.72	1.72	1.72	1.72	1.72	1.72	1.73	1.73
1.73	1.73	1.73	1.74	1.74	1.75	1.75	1.75	1.75	1.75	1.76	1.76
1.76	1.76	1.76	1.76	1.77	1.77	1.77	1.77	1.77	1.77	1.77	1.78
1.78	1.78	1.78	1.78	1.79	1.79	1.79	1.79	1.79	1.79	1.79	1.79
1.80	1.80	1.81	1.81	1.81	1.81	1.81	1.81	1.81	1.82	1.82	1.82
1.83	1.83	1.83	1.84	1.84	1.84	1.85	1.85	1.85	1.85	1.86	1.86
1.86	1.87	1.87	1.88	1.88	1.88	1.88	1.89	1.89	1.90	1.90	1.91

La règle de Sturges donne

$$p = 1 + \log_2(120) \approx 1 + 3.3 \log_{10}(n) \approx \boxed{8}$$

L'amplitude fixe est donc

$$a = (1.92 - 1.60)/8 = \boxed{4}$$

Variable continue: Exemple (suite)

Le tableau des effectifs est

Classes l_i	Effectifs n_i	Fréquences f_i	Effectifs cumulés N_i	Fréquences cumulées F_i
$[1.60, 1.64[$	8	0.07	8	0.07
$[1.64, 1.68[$	14	0.12	22	0.19
$[1.68, 1.72[$	16	0.13	38	0.32
$[1.72, 1.76[$	20	0.17	58	0.49
$[1.76, 1.80[$	26	0.22	84	0.71
$[1.80, 1.84[$	15	0.12	99	0.83
$[1.84, 1.88[$	12	0.10	111	0.93
$[1.88, 1.92[$	9	0.07	120	1
Total	120	1	–	–

N. B. Le recours au regroupement des valeurs en classes peut se faire, aussi, dans le cas discret, lorsque la variable prend un nombre 'assez grand' de valeurs.

Caractère ordinal

Le tableau est dressé suivant la même démarche que dans le cas discret.

Exemple 6 : Une étude sur le niveau de diplôme des 25 managers américains les mieux payés donne (Forbes May 17, 1999):

Top	Noms	Société	Niveau de diplôme
1.	Michael d. Eisner	Walt Disney	Bachelor
2.	Mel Karmazin	CBS	Bachelor
3.	Stephen C. Hilbert	Conseco	None
4.	Millard Drexler	Gap	Master
5.	John F. Welsch, Jr.	General Electric	Doctorat
...			
20.	William R. Steere, Jr.	Pfizer	Bachelor
...			
25.	Richard Jay Kogan	Schering-Plough	Master

Caractère ordinal (suite)

Tableau des effectifs:

Modalités : Diplôme	Effectifs n_i	Fréquences f_i	Effectifs cumulés N_i	Fréquences cumulées F_i
None	1	0.04	1	0.04
Bachelor	7	0.28	8	0.32
Master	11	0.44	19	0.76
Doctorat	6	0.24	25	1
Total	25	1	–	–

Caractère nominal

Le tableau est dressé suivant la même démarche sauf que Les modalités ne sont pas ordonnées.

Exemple 7: Selon *The Nilson Report, Oct. 8, 1998*, les 200 milliards achats par carte de crédits effectués aux USA pendant le premier semestre de l'année 1998 sont répartis selon la marque de la carte utilisée comme suit :

- 36 milliards achats avec la carte American express
- 2 milliards achats avec la carte Diners club
- 12 milliards achats avec la carte Discover
- 50 milliards achats avec la carte Master card
- 100 milliards achats avec la carte Visa.

Caractère nominal (suite)

Tableau des effectifs:

Modalités : Marque	Effectifs en milliard n_i	Fréquences f_i	Pourcentages p_i
American express	36	0.18	18%
Diners club	2	0.01	1%
Discover	12	0.06	6%
Master card	50	0.25	25%
Visa	100	0.50	50%
Total	200	1	100%

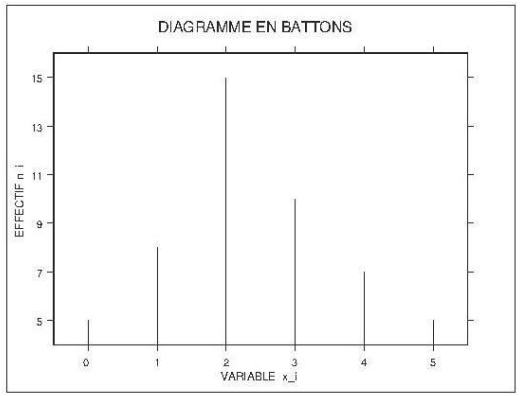
N. B. Les modalités ne sont pas ordonnées, ainsi **la notion de cumule n'a pas de sens**.

Utilité d'un graphe

- Les graphes statistiques sont des figures (images) qui donnent une idée générale sur les données dès le premier coup d'œil ;
- Ils ont l'avantage de la qualité visuel ;
- ils sont utilisées pour
  - resumer des données ;
  - explorer des données ;
  - exposer des résultats ;
  - médiatiser des résultats ;
  - comparer des situations. . .

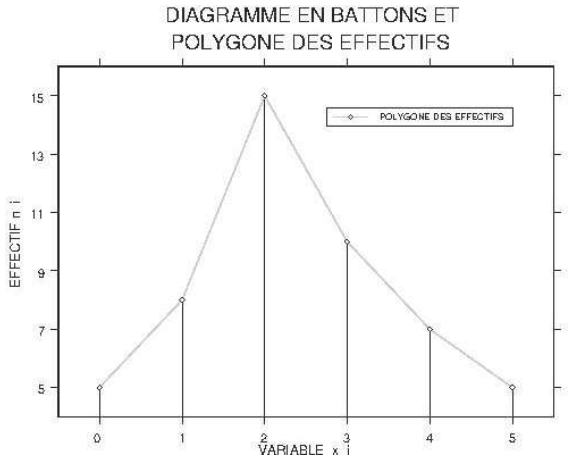
Cas d'une variable discrète: Diagramme en bâtons

Considérons la série discrète de l'Exemple 5: '*les personnes à charge*'. Prenons un repère cartésien orthogonal, et représentons les valeurs de la variable sur l'axe des abscisses et leurs effectifs sur l'axe des ordonnées.



Cas d'une variable discrète: Polygone des effectifs

La liaison des sommets des bâtons par des segments forme le polygone des effectifs.

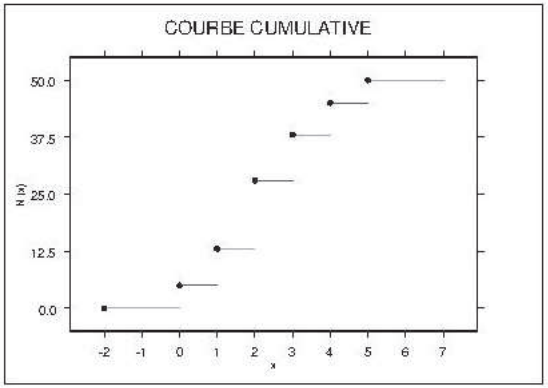


Cas d'une variable discrète: Courbe cumulative

Représentons les effectifs cumulés en définissant une courbe en escalier de la fonction constante par morceaux définie par:

$$N(x) := \begin{cases} 0, & x < x_1; \\ N_1, & x_1 \leq x < x_2; \\ \vdots & \\ N_i, & x_i \leq x < x_{i+1}; \\ \vdots & \\ n, & x_p \leq x. \end{cases}$$

Cas d'une variable discrète: Courbe cumulative (suite)



Cas d'une variable continue

Exemple 8: Le tableau suivant résume les salaires horaires en Dh de 250 employés d'une entreprise:

Classes	Effectifs n_i
$[47, 52[$	10
$[52, 57[$	30
$[57, 60[$	60
$[60, 63[$	72
$[63, 67[$	40
$[67, 77[$	38
Total	250

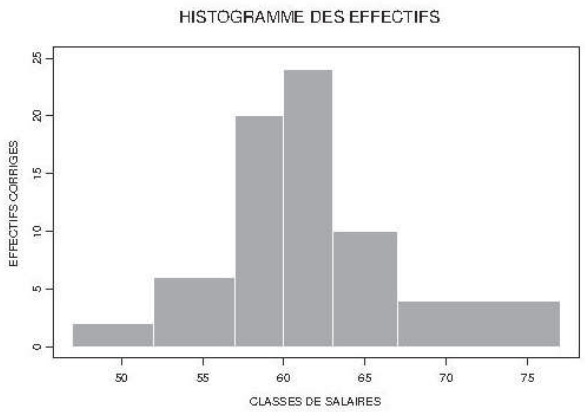
Cas d'une variable continue: Histogramme des effectifs

On représente sur l'axe des abscisses les classes $[x_i^-, x_i^+[$, sur lesquelles un rectangle de surface égale à l'effectif n_i est dressé.

- la base du rectangle $\equiv$ amplitude de la classe $= a_i$;
- l'hauteur du rectangle $\equiv \dfrac{\text{Effectif}}{\text{Amplitude}} = \dfrac{n_i}{a_i}$.
- Le nombre $n_i^* := n_i/a_i$ est appelé **l'effectif corrigé** associé à la classe I_i ; on a corrigé les effectifs.

Classes I_i	Effectifs n_i	Amplitudes a_i	Effectifs corrigés n_i^*	Effectifs cumulés N_i
$[47, 52[$	10	5	2	10
$[52, 57[$	30	5	6	40
$[57, 60[$	60	3	20	100
$[60, 63[$	72	3	24	172
$[63, 67[$	40	4	10	212
$[67, 77[$	38	10	3.8	250
Total	250	30	–	–

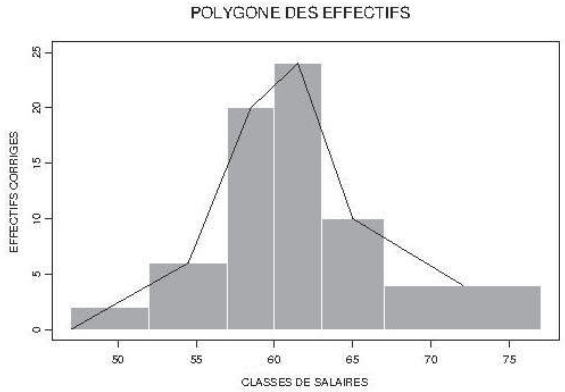
Cas d'une variable continue: Histogramme des effectifs (suite)



N. B. Si les classes sont de mêmes amplitudes, on a pas besoin de corriger les effectifs.

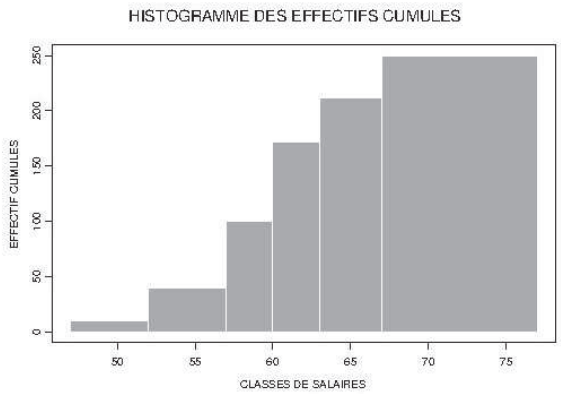
Cas d'une variable continue: Polygone des effectifs

On trace une ligne brisée reliant les points milieu des côtes supérieurs des rectangles



Cas d'une variable continue: Histogramme des effectifs cumulés

On représente sur l'axe des abscisses les classes I_i et sur chaque intervalle $[x_i^-, x_i^+[$ un rectangle de hauteur égale à l'effectif cumulé N_i.

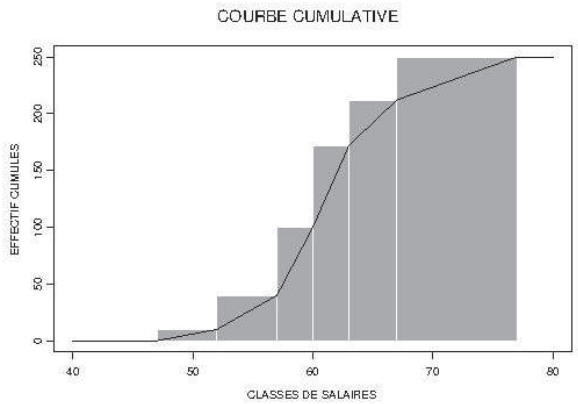


Cas d'une variable continue: Courbe cumulative

On interpole les effectifs cumulés par une ligne brisée joignant les points d'abscisses (x_i^+, N_i).
Plus précisément c'est la courbe représentative de la fonction affine par morceaux définie par:

$$N(x) := \begin{cases} 0, & x < x^-; \\ \dfrac{n_1}{a_1}(x - x_1^-), & x_1^- \leq x < x_1^+; \\ N_1 + \dfrac{n_2}{a_2}(x - x_2^-), & x_2^- \leq x < x_2^+; \\ \vdots & \\ N_{i-1} + \dfrac{n_i}{a_i}(x - x_i^-), & x_i^- \leq x < x_i^+; \\ \vdots & \\ n, & x_p^+ \leq x. \end{cases}$$

Cas d'une variable continue: Courbe cumulative (suite)



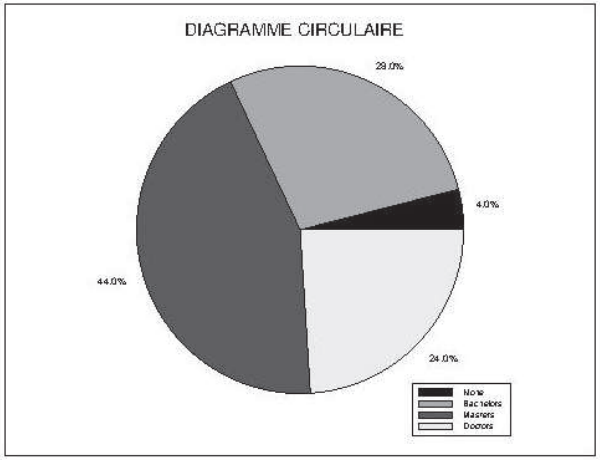
Cas d'un caractère qualitatif: Graphique en secteurs

Reconsidérons l'Exemple 6: '*le niveau de diplôme*'. Dans un cercle on trace des portions (ou secteurs) de surfaces proportionnelles aux effectifs des modalités.
L'angle α_i relatif à la ième modalité est, donc,

$$\alpha_i = 360^\circ \times f_i.$$

Modalités : Diplôme	Fréquences f_i	Angles α_i
None	0.04	14.4°
Bachelor	0.28	100.8°
Master	0.44	158.4°
Doctorat	0.24	86.4°
Total	1	360°

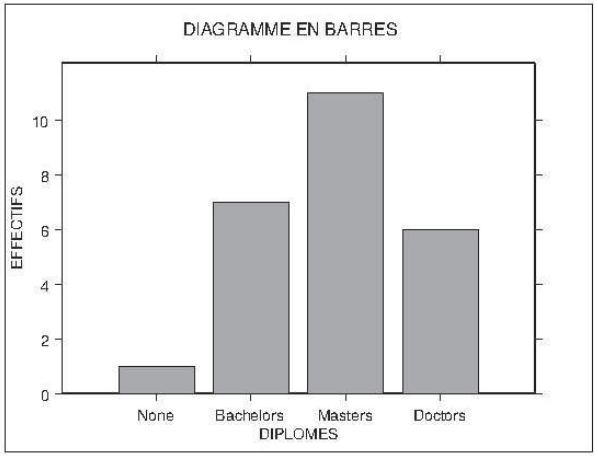
Cas d'un caractère qualitatif: Graphique en secteurs



N. B. Lorsque le nombre de modalités retenues est petit, on peut utiliser un graphe semi-circulaire, dans ce cas : $\alpha_i = 180^\circ \times f_i$.

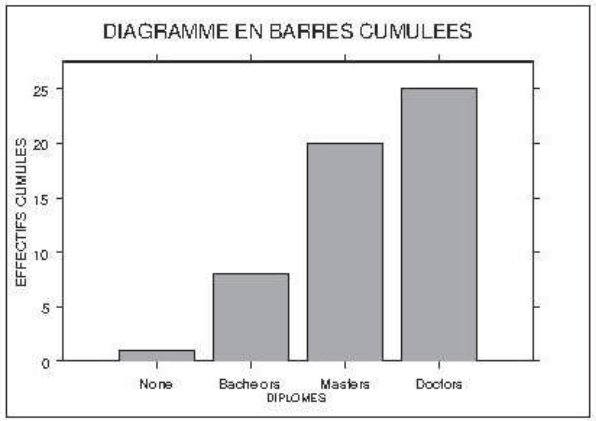
Cas d'un caractère qualitatif: Diagramme en barres

Il s'agit de rectangles de largeurs fixes et d'hauteurs n_i.



Cas d'un caractère ordinal: Diagramme en barres cumulées

Contrairement au cas nominal, dans le cas d'un **caractère ordinale** on peut envisager un diagramme représentant les effectifs cumulés. Il s'agit de rectangles de largeurs fixes et d'hauteurs N_i.



Conclusions

1. Il existe d'autres types de graphiques : Diagramme en tiges (Steam and leaf), Diagramme de Tukey, Cartographie, Pyramide des âges, Graphiques de comparaison, etc.
2. Les graphes sont, donc, assez variés ; chaque graphe s'adapte à une situation particulière.
3. L'essentielle dans une représentation graphique est sa simplicité et sa clarté, il faut donc :
   - inclure toutes les informations utiles à la compréhension du graphique : titre, légende etc ;
   - éviter les informations, mentions et lignes inutiles ;
   - un graphique simple sera préféré à un graphique sophistiqué ;
   - choisir les unités et les axes de la manière la plus neutre possible, il ne faut pas influencer le lecteur ;
   - comparer des graphiques ayant des unités communes.

Introduction

- En pratique on est gêné en présence d'un grand nombre de données
- Si l'intégralité de ces valeurs forme l'information complète, il n'est pas aisé de les manipuler ensemble
- Il faut donc caractériser une variable statistique par un ensemble de paramètres
- Les plus utilisées sont les caractéristiques : de position, de dispersion, de forme
- Les caractéristiques de position nous renseignent sur la position (l'emplacement sur $\mathbb{R}$) de la variable statistique

Mode

Définition

Soit $(x_i, n_i)_{1\leq i\leq p}$ une série statistique de caractère qualitatif ou quantitatif discret.
Le **mode** est la valeur ou la modalité x_j du caractère qui a le plus grand effectif : $\max\limits_{1\leq i\leq p} n_i = n_j$

- **Exemple 5:** Le mode de la série des 'Personnes à charge' est $x_3 = 2$ car $\max\limits_{1\leq i\leq 6} n_i = 15 = n_3$
  Graphiquement, le mode 2 est la valeur de la variable associé au plus grand bâton dans le Diagramme en bâtons
- **Exemple 6:** Le mode de la série des 'Niveaux de diplômes' est la modalité '**Masters**' car $\max\limits_{1\leq i\leq 4} n_i = 11 = n_3$.
  Graphiquement, le mode est la modalité associée au plus large secteur dans le Diagramme en secteurs

Classe modale

Définition

Soit $(l_i, n_i)_{1\leq i\leq p}$ une série statistique continue. La **classe modale** l_j est la classe qui a le plus grand **effectif corrigé** :

$$\max_{1\leq i\leq p} n_i^* = n_j^*$$

- **Exemple 8:** La classe modale de la série des 'Salaires horaires'est la classe $l_4 = [60, 63[$ car $\max\limits_{1\leq i\leq 6} n_i^* = 24 = n_4^*$

  Graphiquement, la classe modale est la modalité associée au plus grand rectangle dans l'Histogramme des effectifs corrigés

Classe modale (suite)

Remarques

1. Lorsque les classe sont de même amplitudes alors la classe modale est la classe qui a le plus grand effectif
   Exemple 2: La classe modale de la série des 'Tailles des sportifs' est $l_5 = [176, 180[$ car $\max\limits_{1\leq i\leq 8} n_i = 26 = n_5$
2. Le mode (la classe modale) n'est pas nécessairement unique : 'Défaut' de cette caractéristique

Moyenne arithmétique

- Soit $\{x_1, \ldots, x_n\}$ des données numériques brutes, alors leurs moyenne arithmétique est

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

- **Exemple:** Soit l'échantillon $\{1, 1, 2, 2, 2, 2, 3, 3\}$ alors

$$\bar{x} = \frac{1}{8}(1 + 1 + 2 + 2 + 2 + 2 + 3 + 3) = 2$$

Les donnés peuvent aussi s'écrire comme série statistique (valeur, effectif): $(1, 2), (2, 4), (3, 2)$ et par suite

$$\bar{x} = \frac{1}{8}(2 \times 1 + 4 \times 2 + 2 \times 3) = 2$$

Moyenne arithmétique (suite)

Définition

Pour une série discrète $(x_1, n_1), (x_2, n_2), \ldots (x_p, n_p)$ la **moyenne arithmétique** est

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_p x_p}{n_1 + n_2 + \ldots + n_p}.$$

Remarque

La formule ci dessus peut être écrite :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i\, x_i \quad \text{ou} \quad \bar{x} = \sum_{i=1}^{p} f_i\, x_i$$

Moyenne arithmétique (suite)

- **Exemple 5:** la moyenne arithmétique de la série des 'personnes à charge' est:

$$\bar{x} = \frac{1}{50} \times (5 \times 0 + 8 \times 1 + 15 \times 2 + 10 \times 3 + 7 \times 4 + 5 \times 5) = \boxed{2.42}$$

  Un salarié a en moyen environ 2.42 personnes à sa charge
- Une série dont les modalités sont qualitatives ne possède pas de moyenne arithmétique

Moyenne arithmétique (suite)

- Pour une série continue $(l_i, n_i)_{1 \leq i \leq p}$, la moyenne arithmétique définie de la même façon en retenant pour x_i les centres des classes l_i
- **Exemple 2:** Tableau de calcul de la moyenne

Classes l_i	Centres x_i	Effectifs n_i	Effectifs $\times$ centres $n_i \times x_i$
[1.60, 1.64[	1.62	8	12.96
[1.64, 1.68[	1.66	14	23.24
[1.68, 1.72[	1.70	16	27.2
[1.72, 1.76[	1.74	20	34.8
[1.76, 1.80[	1.78	26	46.28
[1.80, 1.84[	1.82	15	27.00
[1.84, 1.88[	1.86	12	22.32
[1.88, 1.92[	1.90	9	17.10
Total	–	120	211.02

$$\bar{x} = \frac{211.02}{120} \approx \boxed{1.76 \text{ m}}$$

Propriétés de la moyenne arithmétique

1. La moyenne arithmétique se conserve par changement d'origine et d'unité

- x une variable statistique, x_0 et d deux réels. Soit la variable u issue de x par le changement d'origine et d'unité:

$$u_i = \frac{x_i - x_0}{d} \quad i = 1, \ldots, p$$

- La moyenne arithmétique de u est:

$$\bar{u} = \frac{\bar{x} - x_0}{d}.$$

- L'utilité pratique est la facilité du changement d'unité

Propriétés de la moyenne arithmétique (suite)

2. La moyen arithmétique est associative

- La moyenne globale d'une variable statistique sur l'agrégation de plusieurs populations est la moyenne pondérée des moyennes partielles
- Soit P_a et P_b deux populations d'effectifs n_a et n_b, x une v. s. de moyennes arithmétiques $\bar{x}_a$ et $\bar{x}_b$ sur P_a et P_b respect.
- La moyenne arithmétique de x sur l'agrégation des deux populations $P := P_a \cup P_b$ est

$$\bar{x} = \frac{n_a \bar{x}_a + n_b \bar{x}_b}{n_a + n_b}.$$

- Les intérêts pratiques sont:
  - La possibilité de calculer la moyenne globale sur une population constitué de plusieurs groupes
  - La mise à jour facile de la moyenne dans le cas d'ajout d'une (des) observation(s)

Propriétés de la moyenne arithmétique (suite)

Exemple:

- Le salaire moyen des huits 8 employés d'une entreprise est $\bar{x} = 26585$
- Si l'entreprise recrute deux nouveaux employés qualifiés dont le revenu moyen est 100 000 Dh
- Le nouveau revenu moyen des employés $\bar{x}'$ est

$$\bar{x}' = \frac{8 \times 26585 + 2 \times 100000}{8+2} = \boxed{41\ 268\ \text{Dh}}$$

Propriétés de la moyenne arithmétique (suite)

3. La moyenne arithmétique est sensible à la présence des valeurs aberrantes

- Une valeur **aberrante** est une valeur qui n'est pas du même ordre de grandeur que la plus part des autres observations
- **Exemple:**

$$\begin{aligned} \text{Échantillon } \{1,1,2,2,2,2,3,3\} &\longrightarrow \bar{x} = 2 \\ \text{Échantillon } \{1,1,2,2,2,2,3,\mathbf{300}\} &\longrightarrow \bar{x} = 39.125. \end{aligned}$$

- Cette propriété est en faite un 'défaut' de la moyenne arithmétique Pour y remédier, on peut éliminer 10% des plus grandes et 10% des plus petites valeurs de la variable

Autres moyennes

Définition

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

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

Ainsi, Selon la valeur du paramètre r, on définit plusieurs moyennes :

Autres moyennes (suite)

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

Autres moyennes (suite)

Chacune des moyennes est appropriée à une situation déterminée:

- La Moyenne géométrique pour le calcul de la moyenne des taux de croissance et des indices
  N. B.
  $$\log(\bar{x}_g) = \frac{1}{n}\sum_{i=1}^{p} n_i \log(x_i)$$
  Ainsi, la moyenne géométrique n'est rien d'autre que l'exponentiel de la moyenne arithmétique des logarithmes des données
- La moyenne harmonique est utilisée pour calculer des moyennes de performances unitaires: vitesses moyenne (km/h), consommation (l/km), coût unitaire (Dh/tonne) ...

Médiane

Définition

La **médiane** est une valeur du caractère, notée $x_{1/2}$ (ou Me), partageant une série ordonnée en deux sous ensembles à tailles égales.

Exemple 1: Données brutes Deux situations s'imposent suivant le nombre d'observation n:

- n impairs : $\{3, 5, 7, \boxed{9}, 10, 11, 12\}$. 9 est la valeur médiane, car il y'a autant d'observations inférieures à 9 que d'observations supérieures à 9
- n pairs : $\{3, 5, 7, \boxed{9, 10}, 11, 12, 13\}$. On parle d'un **Intervalle Médian** $[9, 10]$ et on retient le plus souvent comme estimation de la médiane le centre de cet intervalle $x_{1/2} = 9.5$

Médiane d'une série statistique discrète

- S'il existe une valeur x_j telle que $N_j = n/2$ alors

$$x_{1/2} = \frac{x_{j-1} + x_j}{2}$$

- Sinon, la valeur médiane est la plus petite valeur x_j dont l'effectif cumulé dépasse la moitié de l'effectif total:

$$x_{1/2} = x_j \Leftrightarrow N_{j-1} < n/2 \leq N_j.$$

Exemple 2: La grille des salaires pour 100 personnes est

Salaires	Effectifs n_i	Effectifs cumulés N_i
1000	25	25
1800	45	70
2200	30	100
Total	100	–

Le premier effectif cumulé qui dépasse $n/2 = 50$ est $N_2 = 70$
$\Rightarrow x_{1/2} = x_2 = 1800$

Médiane d'une variable continue

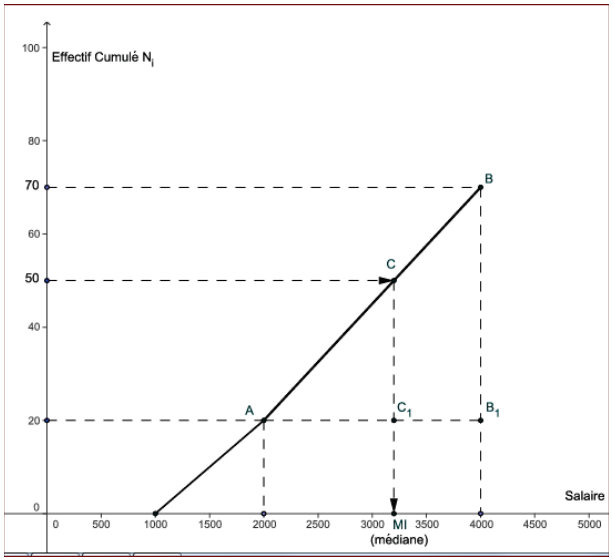
- **Exemple 3:** Considérons par exemple la distribution des salaires observés en continu:

Salaires	Effectifs n_i	Effectifs cumulés N_i
$[1000, 2000[$	20	20
$[2000, 4000[$	50	70
$[4000, 6000[$	30	100
Total	100	–

- La détermination de $x_{1/2}$ se fait en deux étapes:
  1. Détermination de la **classe médiane** $I_m := [x_m^-, x_m^+[$: c'est la première classe dont l'effectif cumulé dépasse $n/2$
     La classe médiane est $I_2 = [2000 - 4000[$; $\Rightarrow x_{1/2} \in I_m = I_2$
  2. Détermination de la valeur médiane par une **interpolation linéaire** suivant la formule suivante:

$$x_{1/2} = x_m^- + (x_m^+ - x_m^-) \times \frac{n/2 - N_{m-1}}{N_m - N_{m-1}} \qquad (1)$$

La méthode d'interpolation



La méthode d'interpolation (suite)

- On trace la courbe cumulative $N(x)$ (une fonction linéaire par morceaux)
- On localise sur l'axe des abscisses la valeur antécédente de la valeur $n/2$ sur l'axe des ordonnées ; cette valeur est la médiane
- On remarque que les deux triangles ABB_1 et ACC_1 sont semblables et par conséquent les rapports des côtés sont égaux:

$$\frac{AC_1}{AB_1} = \frac{CC_1}{BB_1} \Leftrightarrow \frac{x_{1/2} - x_m^-}{x_m^+ - x_m^-} = \frac{n/2 - N_{m-1}}{N_m - N_{m-1}}$$

- d'où la relation (1)
- $\Rightarrow$ la médiane de la série des salaires est

$$x_{1/2} = 2000 + (4000 - 2000) \times \frac{50 - 20}{70 - 20} = \boxed{3\ 200 \text{ Dh}}$$

Quantiles

Définition

Soit $\alpha \in]0,1[$. On appelle **quantile d'ordre** α la valeur x_α de la variable telle que au moins $100 * \alpha\%$ des observations sont inférieures ou égales à x_α .

Remarques

1. Il y'a au moins $100 * (1-\alpha)\%$ des observations qui sont supérieures ou égales à x_α
2. La médiane est un quantile particulier d'ordre $\alpha = 0.5$

Quantiles (suite)

- Empiriquement, les quantiles sont des valeurs qui partagent la série en plusieurs sous-groupes de tailles égaux
- Résumons les quantiles usuels :

Quantiles	Ordres α	Notations	Sous-groupes
Médiane	0.5	Me	$2 \times 50\%$
Quartiles	$(0.25, 0.50, 0.75)$	(Q_1, Q_2, Q_3)	$4 \times 25\%$
Deciles	$(0.10, 0.20, \ldots, 0.90)$	$(D_1, \ldots, D_9)$	$10 \times 10\%$
Centiles	$(0.01, 0.02, \ldots, 0.99)$	$(C_1, \ldots, C_{99})$	$100 \times 1\%$

Quantiles (suite)

La détermination du quantile x_α est identique à celle de la médiane :

1. Détermination de la classe $I_m := [x_m^-, x_m^+[$ qui inclue le quantile: c'est la première classe dont l'effectif cumulé dépasse $\alpha * n$
2. Détermination du quantile d'ordre α par l'interpolation linéaire :

$$x_\alpha = x_m^- + (x_m^+ - x_m^-) \times \frac{\alpha\, n - N_{m-1}}{N_m - N_{m-1}}. \qquad (2)$$

Quantiles (suite)

Déterminons les **quartiles** Q_1, Q_2, Q_3

- Premier quartile $Q_1 = x_{1/4}$. On a $\alpha\, n = 0.25 \times 100 = 25$
  Le premier effectif cumulé dépassant 25 est $N_2 = 70$ ainsi $Q_1 \in I_2 = [2000, 4000[$:

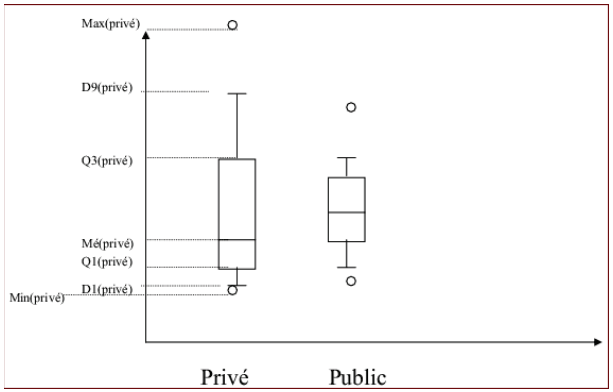
$$Q_1 = 2000 + (4000 - 2000) \times \frac{25 - 20}{70 - 20} = \boxed{2\,200}.$$

- Deuxième quartile $Q_2 = x_{1/2}$. C'est la médiane! $Q_2 = \boxed{3\,200}$
- Troisième quartile $Q_3 = x_{3/4}$. On a $\alpha\, n = 0.75 \times 100 = 75$
  Le premier effectif cumulé dépassant 75 est $N_3 = 100$ ainsi $Q_3 \in I_3 = [4000, 6000[$

$$Q_3 = 4000 + (6000 - 4000) \times \frac{75 - 70}{100 - 70} \approx \boxed{4\,333.33}.$$

Représentation graphique : Boîte à Moustaches

Exemple: Distributions de salaires dans le secteur privé et le secteur public en France :



Boîte à Moustaches (suite)

Définition

La boîte à Moustaches est un rectangle dont : deux côtés sont délimités par les valeurs Q_1 et Q_3.
Deux segments sortent des deux côtés du rectangle et sont délimitées par les valeurs D_1 et D_9
Un segment à l'intérieur du rectangle représente la valeur de la médiane
Deux points en dessous et au dessus de la boîte à moustaches représentent les valeurs Minimum et Maximum

- La **Boîte à Moustaches** (Box Plot), dite aussi **Diagramme de Tukey**, est un graphique synthétique qui permet de représenter : la médiane, les quartiles et les deciles
- il rend compte du niveau d'asymétrie, de la dispersion et des valeurs extrêmes de la série

Conclusions

1. Le tableau ci-dessous résume les paramètres de position envisageable en fonction du type de la variable étudiée :

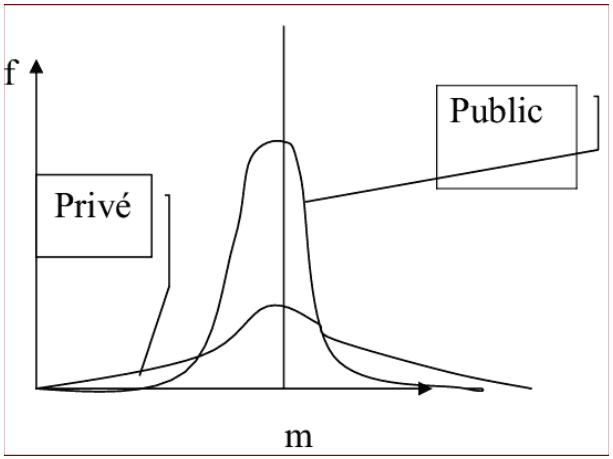
	Quantitative	Ordinale	Nominale
Moyenne	OUI	NON	NON
Médiane	OUI	OUI	NON
Quantiles	OUI	OUI	NON
Modes	OUI	OUI	OUI

2. Le choix d'un paramètre de position dépend de **l'objectif** de l'étude
   - Le lendemain des élections législatives, on cherche à savoir quel est le groupe parlementaire majoritaire, c. à. d. le **mode du vote**
   - La connaissance des **deciles** d'une série des tailles clients est plus utile à une entreprise de confection de vêtement que la connaissance de leurs taille moyenne

Introduction

- On peut remarquer que deux séries statistique peuvent avoir la même valeur centrale mais se différencient par la dispersion des valeurs observées autour de cette dernière
- **Exemple :** Les distributions des salaires des employés dans les secteurs privé et public sont représentées ci-dessous :

Introduction (suite)



Introduction (suite)

- Remarquons que les salaires moyens dans le privé et le public sont équivalents
- Dans le privé il y a une proportion non négligeable qui gagne bien plus que la moyenne et une autre proportion qui gagne bien moins que la moyenne
- Dans le public, en revanche, les salaires sont plus concentrés autour de la moyenne
- On ne peut avoir une idée de la distribution avec seulement la moyenne. Une mesure supplémentaire sur la dispersion autour de cette moyenne doit aussi être donnée
- Tout comme il existe plusieurs valeurs centrales (mode, moyenne, médiane), il existe aussi plusieurs mesures de dispersion.

Étendue

Définition

L'**étendue**, noté E, est la différence entre la plus grande et la plus petite valeur de la variable :

$$E := \max_i (x_i) - \min_i (x_i).$$

- **Exemple:** L'étendue de la série des '*personnes à charge*' est

$$E = 5 - 0 = 5$$

- **Exemple:** L'étendue de la série des '*tailles de sportifs*' est

$$E = 1.91 - 1.60 = 0.31m$$

- L'étendue permet d'avoir une idée de grandeur sur la portée d'une variable statistique

Étendue (suite)

- Cependant il présente le désavantage de ne pas tenir compte de toutes les observations et d'être particulièrement sensible aux valeurs aberrantes
  En effet les deux échantillons

$$E_1 := \{3, 10, 23, 32, 54, 80, 90\} \text{ et } E_2 := \{3, 20, 21, 23, 25, 26, 90\}$$

ont le même étendue : $E = 87$, cependant leurs dispersions sont largement différentes

Écarts inter-quantiles

Définition

Un **écart inter-quantiles** est égal à la différence entre deux quantiles symétriques par rapport à la médiane, c. à. d. pour $0 < \alpha < 1/2$ l'écart est

$$x_{1-\alpha} - x_{\alpha}.$$

- Pour $\alpha = 1/4$, on a un seul **écart interquartile**:

$$EI := x_{3/4} - x_{1/4} = Q_3 - Q_1$$

- pour $\alpha = 1/10, 2/10, 3/10, 4/10$, on a quatre **écarts inter-deciles**:

$$D_9 - D_1, \quad D_8 - D_2, \quad D_7 - D_3, \quad D_6 - D_4$$

Écarts inter-quantiles (suite)

- Dans un Diagramme en Boites à Moustaches on peut lire les écarts, particulièrement importants, $Q3 - Q1$ et $D9 - D1$
- Ils présentent une mesure alternative à l'étendue, puisqu'ils ne tiennent pas compte des valeurs maximales et minimales pouvant être aberrantes

Remarque

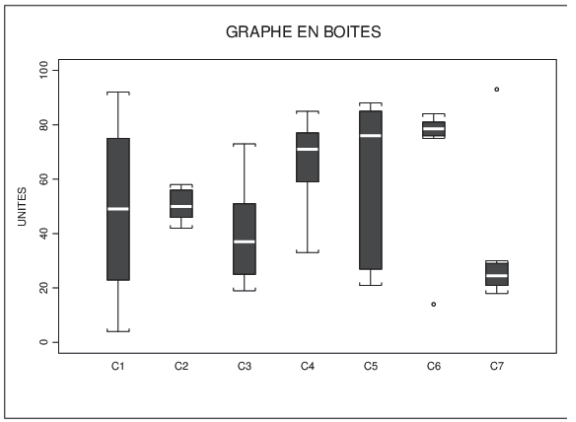
L'écart interquartile est généralement utilisé pour détecter l'existence éventuelle des valeurs aberrantes
Plusieurs logiciels comme le **SPSS**®, considèrent comme valeur aberrante toute valeur x^* qui se trouve à l'extérieur de l'intervalle $[Q_1 - 1.5 * EI, Q_3 + 1.5 * EI]$.

Écarts inter-quantiles (suite)

Exemple : Les résultats (sur 100) de 10 étudiants pour 7 cours sont présentés dans le tableau suivant :

C1	C2	C3	C4	C5	C6	C7
4.00	42.00	19.00	33.00	21.00	14.00	18.00
12.00	44.00	23.00	47.00	25.00	75.00	19.00
23.00	46.00	25.00	59.00	27.00	76.00	21.00
35.00	47.00	27.00	67.00	29.00	77.00	23.00
46.00	49.00	31.00	69.00	77.00	78.00	24.00
52.00	51.00	43.00	73.00	75.00	79.00	25.00
67.00	54.00	48.00	75.00	83.00	80.00	27.00
75.00	56.00	51.00	77.00	85.00	81.00	29.00
83.00	57.00	63.00	83.00	88.00	83.00	30.00
92.00	58.00	73.00	85.00	87.00	84.00	93.00

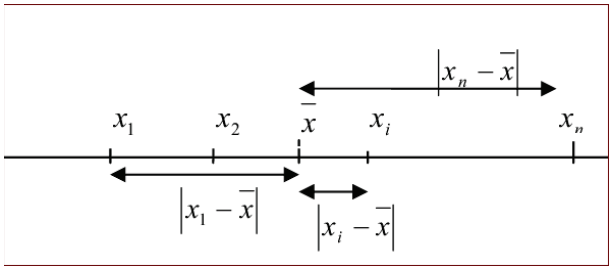
Écarts inter-quantiles (suite)



Classement par ordre croissant des 7 cours selon la dispersion des résultats: C6, C7, C2, C4, C3, C5 et C1

Écart moyen absolu

- Considérons une valeur centrale comme par exemple la moyenne $\bar{x}$
- Une forte concentration des observations autour de $\bar{x}$ se traduit par des écarts $|x_i - \bar{x}|$ faibles et inversement
- Une façon de mesurer la dispersion est, donc, le calcul de la moyenne de ces écarts



Écart moyen absolu (suite)

Définition

Soit $(x_i, n_i)_{1 \leq i \leq p}$ une série statistique quantitative.

1. L'**écart moyen absolu**, noté e_m, est défini par :

$$e_m := \frac{1}{n} \sum_{i=1}^{p} n_i \, |x_i - \bar{x}|.$$

2. L'**écart médian absolu**, noté e_m^*, est défini par :

$$e_m^* := \frac{1}{n} \sum_{i=1}^{p} n_i \, |x_i - x_{1/2}|.$$

N. B. Ces écarts représentent l'inconvenient de n'être pas analytique à cause de la valeur absolue : fonction non derivable et peu commode pour les calculs

Variance et Écart-type

La remplacement de la valeur absolue par le carré, donne lieu à une valeur typique de la dispersion dite la **variance**

Définition

La **variance**, notée S^2, d'une variable statistique est définie par

$$S^2 := \frac{1}{n}\sum_{i=1}^{p} n_i\,(x_i - \bar{x})^2.$$

Contrairement à l'étendue et aux quartiles, la variance permet de combiner toutes les valeurs à l'intérieur d'un ensemble de données afin d'obtenir la mesure de dispersion.

Variance et Écart-type (suite)

Théorème de Konig Huyghens

La variance est égale à la moyenne des carrés moins le carré de la moyenne.

$$S^2 = \left(\frac{1}{n}\sum_{i=1}^{p} n_i\,x_i^2\right) - \bar{x}^2.$$

Preuve : Soit $(x_i, n_i)_{1\leq i\leq p}$ une série statistique quantitative.

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{p} n_i\,(x_i - \bar{x})^2. \\ &= \frac{1}{n}\sum_{i=1}^{p} n_i\,(x_i^2 - 2\,x_i\,\bar{x} + \bar{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{p} n_i\,x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{p} n_i x_i + \bar{x}^2\frac{1}{n}\sum_{i=1}^{p} n_i \end{aligned}$$

Variance et Écart-type (suite)

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

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{p} n_i\, x_i^2 - 2\bar{x}^2 + \bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{p} n_i\, x_i^2 - \bar{x}^2. \end{aligned}$$

□

Remarque

Il est plus facile d'appliquer la formule de *Konig Huyghens* ; Elle permet d'éviter la cumulation des erreurs d'arrondis

Variance et Écart-type (suite)

- La variance S^2 s'exprime en carré de l'unité de la variable
- Afin de revenir à l'unité de la variable on définit L'**écart-type**, noté S, comme étant la racine carrée de la variance :

$$S := \sqrt{S^2}.$$

Variance et Écart-type (suite)

Exemple: L'écart-type de la série des '*Tailles des sportifs*' peut être obtenue en utilisant le tableau suivant:

l_i	x_i	n_i	$n_i \times x_i$	$n_i \times x_i^2$
$[1.60, 1.64[$	1.62	8	12.96	21.00
$[1.64, 1.68[$	1.66	14	23.24	38.58
$[1.68, 1.72[$	1.70	16	27.2	46.24
$[1.72, 1.76[$	1.74	20	34.8	60.55
$[1.76, 1.80[$	1.78	26	46.28	82.38
$[1.80, 1.84[$	1.82	15	27.00	49.69
$[1.84, 1.88[$	1.86	12	22.32	41.52
$[1.88, 1.92[$	1.90	9	17.10	32.49
Total	–	120	211.02	372.43

Ainsi, $\bar{x} = 211.02/120 \approx 1.76$, $S^2 = (372.43/120) - 1.76^2 \approx 0.0006$ et

$$S = \sqrt{0.0006} \approx \boxed{0.078 \text{ m}}$$

Propriétés de l'écart type

1. **L'écart-type ainsi que la variance ne sont jamais négatif.** Ils sont nulles si et seulement si toutes les valeurs de la variable sont égales, c. à. d. quand il y'a pas de variation
2. **L'écart-type est sensible aux valeurs aberrantes.** Une seule valeur aberrante peut accroître l'écart-type et, par le fait même, déformer le portrait de la dispersion. Il peut être donc un bon indicateur aussi de valeurs aberrantes
3. **L'écart-type est invariant par changement d'origine.** Soit x une variable statistique, x_0 et d deux réels et u est la variable issue de x par le changement d'origine et d'unité suivant:

$$u_i = \frac{x_i - x_0}{d} \quad i = 1, \ldots, p.$$

Alors, la variance et l'écart-type de u sont respectivement

$$S_u^2 = \frac{S_x^2}{d^2} \quad \text{et} \quad S_u = \frac{S_x}{|d|}.$$

Coefficient de variation

- Lorsque les moyennes ne sont pas de même ordre de grandeur ou lorsque ils n'ont pas les même unités, pour comparer les écarts types il faut les exprimer par rapport à la moyenne
- Il faut comparer des écarts types unitaires d'où l'introduction du **coefficient de variation**, noté CV,

$$CV := \frac{S}{\bar{x}}$$

- Le coefficient de variation est une mesure de la dispersion relative; il est **adimensionnel** (nombre sans unité) : on peut l'exprimé en pourcentage

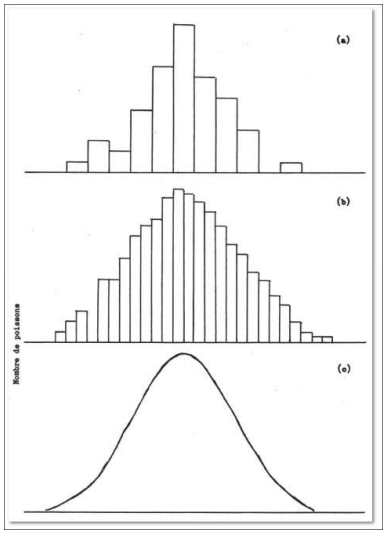
Coefficient de variation (suite)

Exemple : La demande d'importation sur une période de 30 ans est résumée dans le tableau ci dessous :

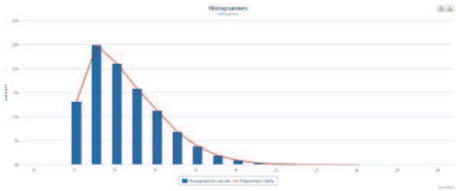
	France (euro)	Allemagne (euro)	Canada ($ Can.)	États Unis (euros conv)
Moyenne	5	5	6	100
Écart type	2.5	1	3	10
CV	50%	20%	50%	10%

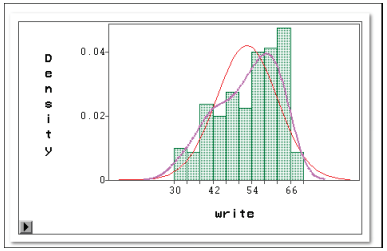
- Comparaison entre séries France et Allemagne : Même moyenne, même monnaie : la dispersion est plus grande pour la France
- Comparaison entre séries France et U.S : écart type des US (monnaie déjà convertit en euros) est plus grand. Mais, la demande moyenne d'importation est beaucoup plus grande aussi pour les États Unis
- Afin de comparer ces dispersions il faut, donc, utiliser les CV ; on voit que la dispersion relative des États Unis est la plus faible

Quelques formes des distributions statistiques



Quelques formes des distributions statistiques (suite)



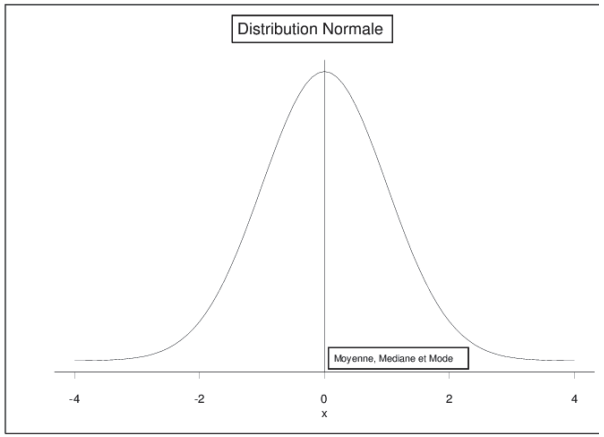


Une distribution particulière : la normale

- Afin de caractériser la forme d'une distribution d'une série, sans avoir à la tracer pour autant, nous choisissons distribution de 'référence' dite distribution **Normale** ou **Gaussienne**
- Il s'agit d'une distribution très populaire introduite par *Laplace* pour approcher la répartition des erreurs de mesures
- C'est une distribution symétrique sous forme de cloche, avec un unique mode (uni modale), dont la moyenne est égale à la médiane et au mode ($\bar{x} = x_{1/2} = Mode$)
- La fonction de densité d'une normale standard est

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

La forme d'une distribution normale



Moments Centrés

- Afin de distinguer les formes des distributions statistiques on utilise des statistiques dites **moments centrés**

Définition

Soient $(x_i, n_i)_{1 \leq i \leq p}$ une série statistique et $r \in \mathbb{Q}$. Le **moment centré d'ordre r** s'écrit :

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

- Pour $r = 1$, le moment centré d'ordre 1 est nul : $m_1 = 0$
- Pour $r = 2$, le moment centré d'ordre 2 n'est rien d'autre que la variance $m_2 = S^2$

Symétrie et Asymétrie

On distingue 3 types de distribution selon leurs asymétries par rapport à la moyenne

1. Distribution symétrique par rapport à la moyenne: Figure 1
2. Distribution dissymétrique à gauche : Figure 2
3. Distribution dissymétrique à droite : Figure 3

Figure 1: Distribution symétrique

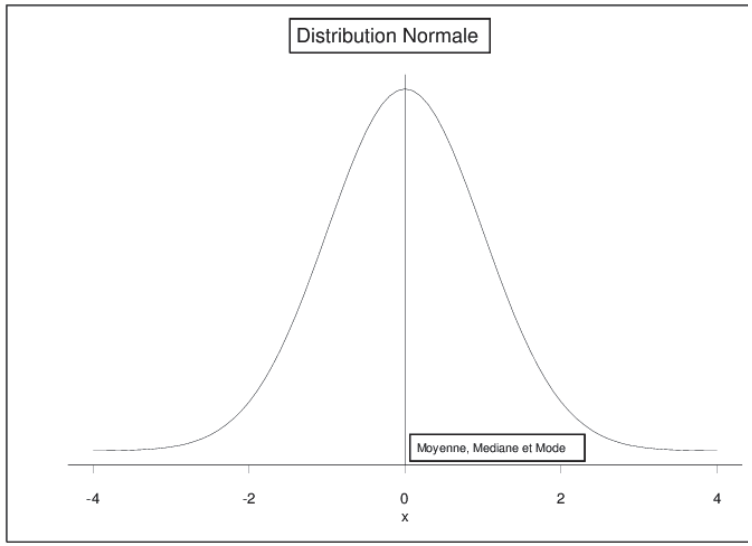


Figure 2: Distribution dissymétrique à gauche

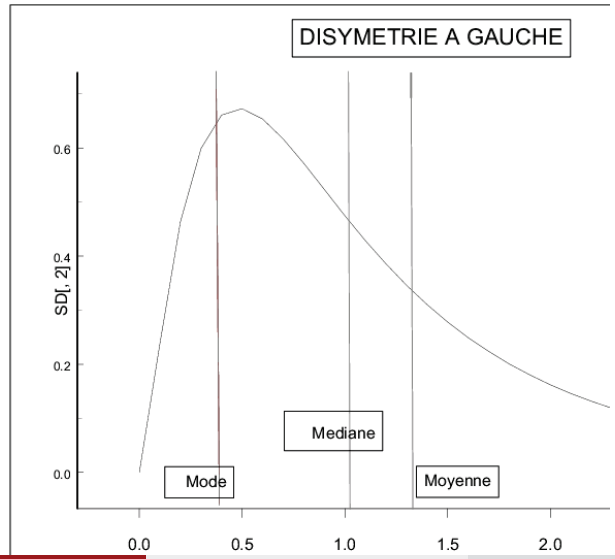
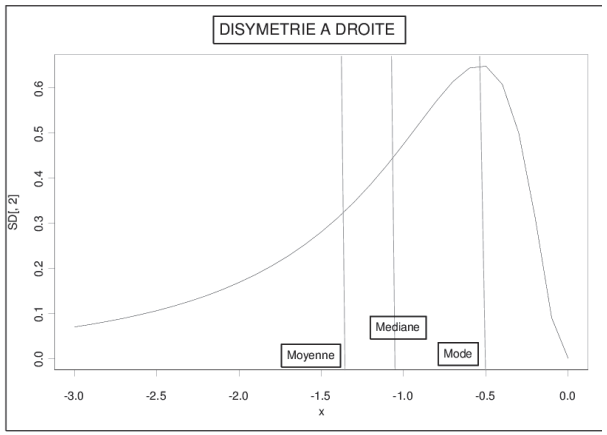


Figure 3: Distribution dissymétrique à droite



Caractérisation de l'asymétrie

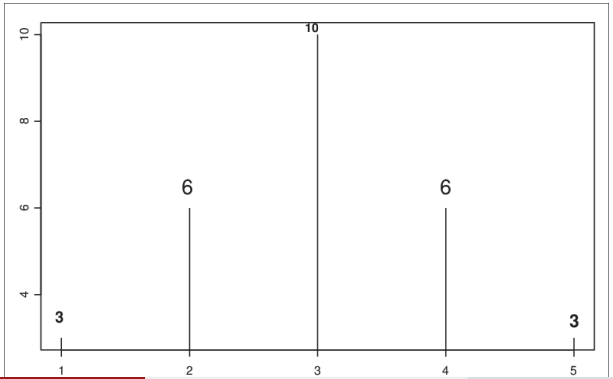
- Le paramètre qui caractérise l'asymétrie d'une distribution $(x_i, n_i)_{1 \leq i \leq p}$ est m_3 le moment centré d'ordre 3
- On constate que pour une distribution

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

Caractérisation de l'asymétrie (suite)

Exemple: Calcul de m_3 de la distribution ci dessous

$$\begin{aligned} m_3 &= \frac{1}{28}[3(1-3)^3 + 6(2-3)^3 + 10(3-3)^3 + 6(4-3)^3 + 3(5-3)^3] \\ &= \frac{1}{28}[-24 + -6 + 0 + 6 + 24] = 0. \end{aligned}$$



Coefficient d'asymétrie de Fisher

- Le moment m_3 est invariant par un changement d'origine; mais il dépend des unités choisies
- Pour cela *Fisher* a introduit un coefficient adimensionnel

Définition

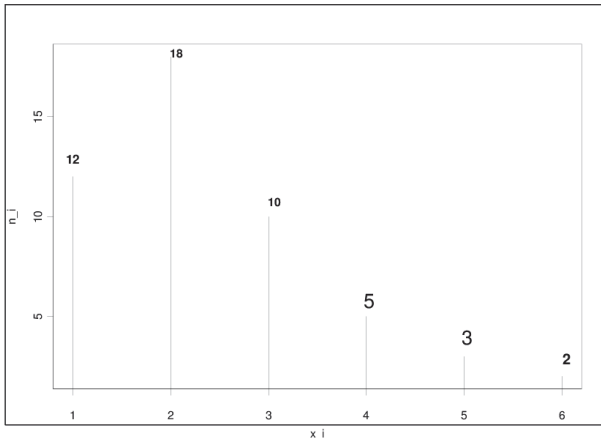
Le **Coefficient d'asymétrie de Fisher (Skewness)** est définit par

$$\gamma := \frac{m_3}{S^3},$$

où m_3 est le moment centré d'ordre 3.

Coefficient d'asymétrie de Fisher : Exemple

Exemple: Considérons la distribution ci dessous



Calcul du coefficient d'asymétrie de Fisher

x_i	n_i	$n_i\,x_i$	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$n_i\,(x_i-\bar{x})^2$	$n_i\,(x_i-\bar{x})^3$
1	12	12	-1.5	2.25	27	-40.500
2	18	36	-0.5	0.25	4.5	-2.250
3	10	30	0.5	0.25	2.5	1.250
4	5	20	1.5	2.25	11.25	16.875
5	3	15	2.5	6.25	18.75	46.875
6	2	12	3.5	12.25	24.5	85.750
$\sum$	50	125	–	–	88.5	$108 > 0$

Grâce au Tableau ci dessus on trouve :

$$\bar{x} = 125/50 = 2.5 \quad ; \quad S^3 = \left(\sqrt{88.5/50}\right)^3 \approx 2.35$$
$$m_3 = 108/50 = \boxed{2.16} \quad ; \quad \gamma \approx 2.16/2.35 \approx 0.92 > 0$$

En conclusion, la distribution est dissymétrique à gauche

Mesure de l'aplatissement

- L'aplatissement d'une distribution est basé sur le m_4: moment centré d'ordre 4

Définition

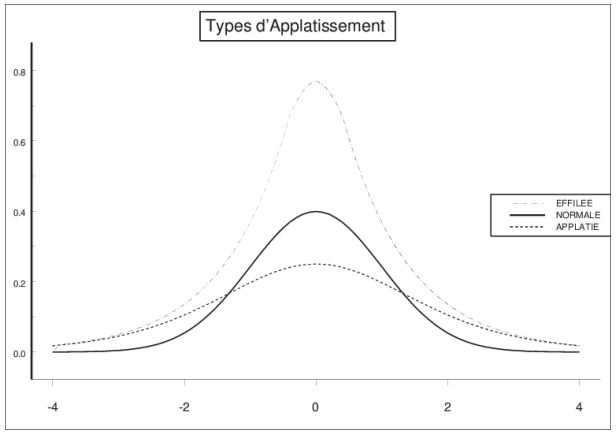
(i) Le **Coefficient d'aplatissement de Pearson** : $\eta := \dfrac{m_4}{S^4}$

(ii) le **Coefficient d'aplatissement de Fisher (Kurtosis)** : $\beta := \dfrac{m_4}{S^4} - 3.$

- Pour la distribution normale : $\eta = 3$ et $\beta = 0$
- Prenons l'aplatissement d'une normale comme 'reference', ainsi une distribution est :

$$\begin{cases} \text{aplatie} & : \ \beta < 0 \ ; \\ \text{normale} & : \ \beta = 0 \ ; \\ \text{effilée} & : \ \beta > 0. \end{cases}$$

Mesure de l'aplatissement: Exemple



Courbe de Lorenz

- La variance ou le coefficient de variation mesurent la concentration des valeurs autour de la moyenne
- Lorsque la distribution est asymétrique, ces mesures ne sont pas entièrement satisfaisantes. Les statisticiens ont alors développé des indices spécifiques pour mesurer la concentration
- Prenons le cas de la distribution des revenus. C'est une distribution asymétrique avec peu de personnes dans les classes de revenus élevés
- En 1905 *Max Otto Lorenz* a proposé un graphique spécial pour représenter la concentration des richesses :
  - en abscisse on met le pourcentage cumulé de la population associée à la classe i
  - en ordonnée le pourcentage cumulée des richesses de la classe i

Courbe de Lorenz: Exemple

Exemple: Le revenu journalier de 10 employés est distribué selon le tableau suivant :

Revenu Classe I_i	Effectifs n_i
$[90, 150[$	5
$[150, 250[$	3
$[250, 550[$	2
Total	10

Courbe de Lorenz: Exemple (suite)

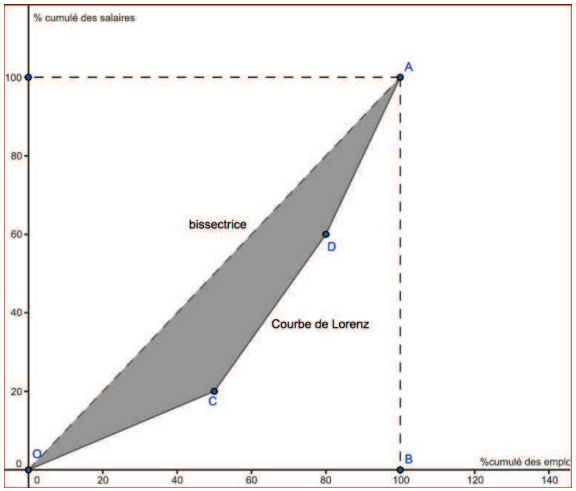
Calculons les pourcentages cumulés des employés et ceux cumulés des salaires :

1. p_i est le pourcentage de I_i, et $\boldsymbol{P_i}$ son cumule
2. $n_i \times x_i$ est la masse salariale attribuée au individus de la classe I_i
3. $\sum_i n_i \times x_i = 2000$ est la masse salariale totale
4. $q_i := 100 \times n_i.x_i / \sum_i x_i.n_i$ est le pourcentage de la masse salariale attribuée au individus de la classe I_i
5. $\boldsymbol{Q_i}$ le cumule de q_i

Classe I_i	x_i	n_i	p_i	P_i	$n_i \times x_i$	q_i	Q_i
$[90, 150[$	120	5	50%	50%	600	30%	30%
$[150, 250[$	200	3	30%	80%	600	30%	60%
$[250, 550[$	400	2	20%	100%	800	40%	100%
Total	–	10	100%	–	2000	100%	–

Courbe de Lorenz (suite)

La **courbe de Lorenz** est la courbe représentative du Q_i en fonction du P_i : c'est le polygone joignant les points $M_i(P_i, Q_i)$



Courbe de Lorenz (suite)

- Si touts les individus percevaient le même salaire, on obtiendrait la première bissectrice : ligne d'égalité parfaite (impossible en réalité!!)
- La surface entre la bissectrice et la courbe représente (surface colorée) est la **surface de concentration**
- Lorsque la répartition devient plus inégale la surface de concentration augmente

Indice de Gini

- En 1912, le statisticien *Corrado Gini* a proposé un indice, noté I, pour mesurer le degré d'inégalité dans la distribution des richesses
- Il est calculé en prenant le rapport entre S_{OADC} la surface du polygone $OADC$ et S_{OAB} la surface du triangle OAB :

$$I := \frac{\text{Aire de concentration}}{\text{Aire du triangle OAB}} = \frac{S_{OADC}}{10^4/2}.$$

- Il est évident que $0 \leq I \leq 1$
- Si la distribution est égalitaire, la courbe de Lorenz suit la diagonale et alors l'indice de Gini est égal à 0
- L'inégalité la plus forte est obtenue lorsque la courbe de Lorenz est OAB ; dans ce cas P_i accroît beaucoup plus vite que Q_i et l'indice de Gini est égal à 1

Indice de Gini (suite)

- La formule analytique de l'**indice de Gini** est

$$I = 1 - 10^{-4} \sum_{i=1}^{p} (P_i - P_{i-1})(Q_{i-1} + Q_i)$$

où $P_0 = Q_0 = 0$

- Calcul du coefficient de Gini dans l'exemple

Classe I_i	x_i	n_i	p_i	P_i	$n_i \times x_i$	q_i	Q_i
$[90, 150[$	120	5	50%	50%	600	30%	30%
$[150, 250[$	200	3	30%	80%	600	30%	60%
$[250, 550[$	400	2	20%	100%	800	40%	100%
Total	–	10	100%	–	2000	100%	–

$$I = 1 - \frac{[50 * 30 + (80 - 50) * (30 + 60) + (100 - 80) * (100 + 60)]}{10^4} = \mathbf{0.26}$$

- L'indice de concentration de Gini n'a de sens que pour les variables cumulable : richesse, revenu, salaire, consommation....

La médiale

Définition

La **médiale** est la valeur du caractère pour laquelle les individus dont le caractère est inférieur à la médiale se partagent 50% de la masse totale des richesses.

- **Rappel:** La médiane est la valeur du caractère pour laquelle les individus dont le caractère est inférieur à la médiane forment 50% de la masse population
- La position de la médiale par rapport à la médiane est un indicateur de concentration

La médiale (suite)

- Si la médiane est inférieure à la médiale ($Me < MI$) alors 50% de la population partagent moins de 50% de la masse salariale
- Tandis que dans une situation imaginaire ou tous les individus percevraient un salaire identique alors la médiane serait égale à la médiale ($Me = MI$)
- Par conséquent plus l'écart entre la médiane et la médiale est important plus la distribution est inégalitaire
- La détermination de MI se fait en deux étapes :
  1. Détermination de la **classe médiale** : celle qui inclue la médiale. La classe médiale $I_m := [x_m^-, x_m^+[$ c'est la première classe dont le cumule Q_i dépasse 50%
  2. Détermination de la valeur médiale par une interpolation linéaire:

$$MI = x_m^- + (x_m^+ - x_m^-) \times \frac{50\% - Q_{m-1}}{Q_m - Q_{m-1}}.$$

La médiale (suite)

La détermination de MI de la série de l'exemple précédent

1. la classe médiale est $I_2 := [150, 250[$ car Q_2 est le premier qui a dépassé 50%
2. l'interpolation linéaire donne

$$MI = 150 + (250 - 150) \times \frac{50 - 30}{60 - 30} \approx 217 Dh$$

La médiane $Me = 150$ Dh, ainsi $MI - ME = 217 - 150 = 67 > 0$ distribution des salaires est non égalitaires.

La médiale (suite)