Cours réalisé par Laurent DOYEN

La statistique descriptive

1. Introduction et définitions

Statistique descriptive:

Analyse et synthèse, **NUMERIQUE** et **GRAPHIQUE**, d’un ensemble de données

1. Introduction et définitions

Statistique descriptive:

Analyse et synthèse, **NUMERIQUE** et **GRAPHIQUE**, d'un ensemble de données

↓

But: Synthétiser l'information contenue dans les données

Origine: étude démographique

Individus: latin: « ce qui est indivisible »
stat: chacune des « personnes » étudiées

Individus: latin: « ce qui est indivisible »
stat: chacune des « personnes » étudiées

Personne humaine, automobile, entreprise, pays, ….

Individus: latin: « ce qui est indivisible »
stat: chacune des « personnes » étudiées

Personne humaine, automobile, entreprise, pays, ….

Population: ensemble des individus observés

Individus: latin: « ce qui est indivisible »
stat: chacune des « personnes » étudiées

Personne humaine, automobile, entreprise, pays, ….

Population: ensemble des individus observés

Les étudiants de 12-25ans, les Renault produites entre 1990 et 1995

Individus: latin: « ce qui est indivisible »
stat: chacune des « personnes » étudiées

Personne humaine, automobile, entreprise, pays, ….

Population: ensemble des individus observés

Les étudiants de 12-25ans, les Renault produites entre 1990 et 1995

Caractère (Variable Statistique): ce qu’on observe sur chacun des individus de la population

Individus: latin: « ce qui est indivisible »
stat: chacun des « personnes » étudiées

Personne humaine, automobile, entreprise, pays, ….

Population: ensemble des individus observés

Les étudiants de 12-25ans, les Renault produites entre 1990 et 1995

Caractère (Variable Statistique): ce qu'on observe sur chacun des individus de la population

Sexe, age, taille, nombre enfants,…

Attention:

La **population** doit être **définie avec précision**, c’est totalement différent de considérer:

- Les étudiants
- Les étudiants de 12-25 ans
- Les étudiants de l’IUP com. et vente de Grenoble

Attention:

La **population** doit être **définie avec précision**,
c'est totalement différent de considérer:

- Les étudiants
- Les étudiants de 12-25 ans
- Les étudiants de l'IUP com. et vente de Grenoble

La **population** doit être **homogène** au regard des caractères étudiés:
la répartition des individus selon leur taille doit distinguer les deux sexes

2 types de caractères:

Qualitatifs: non mesurables

2 types de caractères:

Qualitatifs: non mesurables

Sexe, couleur des yeux, secteur d'activité

<u>2 types de caractères:</u>

Qualitatifs: non mesurables

Sexe, couleur des yeux, secteur d'activité

Quantitatifs: mesurables

2 types de caractères:

Qualitatifs: non mesurables

Sexe, couleur des yeux, secteur d'activité

Quantitatifs: mesurables

Age, taille , PIB, taux de chômage

2 types de caractères:

Qualitatifs: non mesurables

Sexe, couleur des yeux, secteur d'activité

Quantitatifs: mesurables

Age, taille , PIB, taux de chômage

Quantitatifs discrets: peuvent prendre un nombre fini et faible de valeurs

2 types de caractères:

Qualitatifs: non mesurables

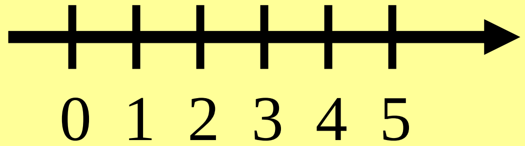
Sexe, couleur des yeux, secteur d’activité

Quantitatifs: mesurables

Age, taille , PIB, taux de chômage

Quantitatifs discrets: peuvent prendre un nombre fini et faible de valeurs

Nb enfants



2 types de caractères:

Qualitatifs: non mesurables

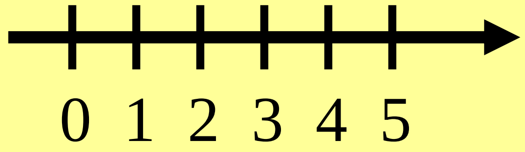
Sexe, couleur des yeux, secteur d'activité

Quantitatifs: mesurables

Age, taille , PIB, taux de chômage

Quantitatifs discrets: peuvent prendre un nombre fini et faible de valeurs

Nb enfants



Quantitatifs continues:

- Par nature:

2 types de caractères:

Qualitatifs: non mesurables

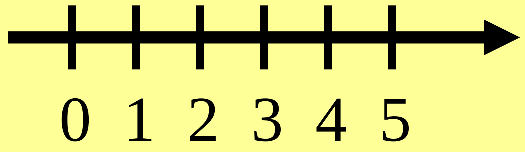
Sexe, couleur des yeux, secteur d'activité

Quantitatifs: mesurables

Age, taille , PIB, taux de chômage

Quantitatifs discrets: peuvent prendre un nombre fini et faible de valeurs

Nb enfants



Quantitatifs continues:

•Par nature:

Taille:

2 types de caractères:

Qualitatifs: non mesurables

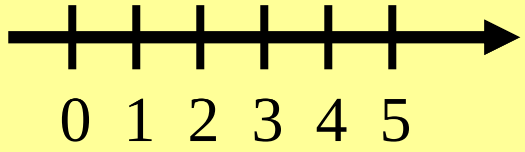
Sexe, couleur des yeux, secteur d’activité

Quantitatifs: mesurables

Age, taille , PIB, taux de chômage

Quantitatifs discrets: peuvent prendre un nombre fini et faible de valeurs

Nb enfants



Quantitatifs continues:

- Par nature:

Taille:

- Par nécessité:

2 types de caractères:

Qualitatifs: non mesurables

Sexe, couleur des yeux, secteur d'activité

Quantitatifs: mesurables

Age, taille , PIB, taux de chômage

Quantitatifs discrets: peuvent prendre un nombre fini et faible de valeurs

Nb enfants

0 1 2 3 4 5

Quantitatifs continues:

•Par nature:

Taille:

1m 2m

•Par nécessité:

Nombre de salariés d'une PME

0 500

2. Étude d'un caractère qualitatif

2.1 Modalités d'un caractère: les différents états d'un caractère qualitatif.

EXHAUSTIFS et INCOMPATIBLES

2. Étude d'un caractère qualitatif

2.1 Modalités d'un caractère: les différents états d'un caractère qualitatif.

EXHAUSTIFS et INCOMPATIBLES

Cad chaque individu présente une et une seule modalité du caractère

2. Étude d'un caractère qualitatif

2.1 Modalités d'un caractère: les différents états d'un caractère qualitatif.

EXHAUSTIFS et INCOMPATIBLES

Cad chaque individu présente une et une seule modalité du caractère

Cadre supérieure, Profession int., Employé, Ouvrier, Ouvrier qualifié

2. Étude d'un caractère qualitatif

2.1 Modalités d'un caractère: les différents états d'un caractère qualitatif.

EXHAUSTIFS et INCOMPATIBLES

Cad chaque individu présente une et une seule modalité du caractère

Cadre supérieure, Profession int., Employé, Ouvrier, Ouvrier qualifié
Inactifs

2. Étude d’un caractère qualitatif

2.1 Modalités d’un caractère: les différents états d’un caractère qualitatif.

EXHAUSTIFS et INCOMPATIBLES

Cad chaque individu présente une et une seule modalité du caractère

Cadre supérieure, Profession int., Employé, Ouvrier, ~~Ouvrier qualifié~~

Inactifs

N= Effectif total de la population

n_i= Effectif de la modalité considérée

2.2 $\underbrace{\text{Pourcentage}}_{p_i}$ et $\underbrace{\text{fréquence}}_{f_i}$:

$$p_i = \frac{n_i}{N} 100 \qquad f_i = \frac{n_i}{N}$$

N= Effectif total de la population

n_i= Effectif de la modalité considérée

2.2 $\underbrace{\text{Pourcentage}}_{p_i}$ et $\underbrace{\text{fréquence}}_{f_i}$:

$$p_i = \frac{n_i}{N} 100 \qquad f_i = \frac{n_i}{N}$$

<u>Propriété:</u> $\sum_i p_i = 100 \qquad \sum_i f_i = 1$

2.2 $\underbrace{\text{Pourcentage}}_{p_i}$ et $\underbrace{\text{fréquence}}_{f_i}$:

N= Effectif total de la population

n_i= Effectif de la modalité considérée

$$p_i = \frac{n_i}{N} 100 \qquad f_i = \frac{n_i}{N}$$

<u>Propriété:</u> $\sum_i p_i = 100 \qquad \sum_i f_i = 1$

<u>Exemple:</u> En 1989 parmi les français de plus de 15 ans

Sur 21033906 hommes il y a 4286858 retraités

$\frac{4286858}{21033906} 100 \approx 20\%$ des hommes sont retraités

2.3 Tableau de distribution:

Français de plus de 15 ans en 1986

CSP	Nb de personnes	Pourcentages
Agriculteurs exploitants	1268264	2.9
Artisans, commerçants et chefs d'entreprises	1757221	4.0
Cadres et professions intellectuelles supérieures	2314770	5.3
Professions intermédiaires	4593294	10.4
Employés	6771239	15.4
Ouvriers	7121812	16.2
Retraités	8429509	19.2
Inactifs divers (autres que retraités)	11741884	26.7
Ensemble	43997993	100

2.4 Représentations graphiques:

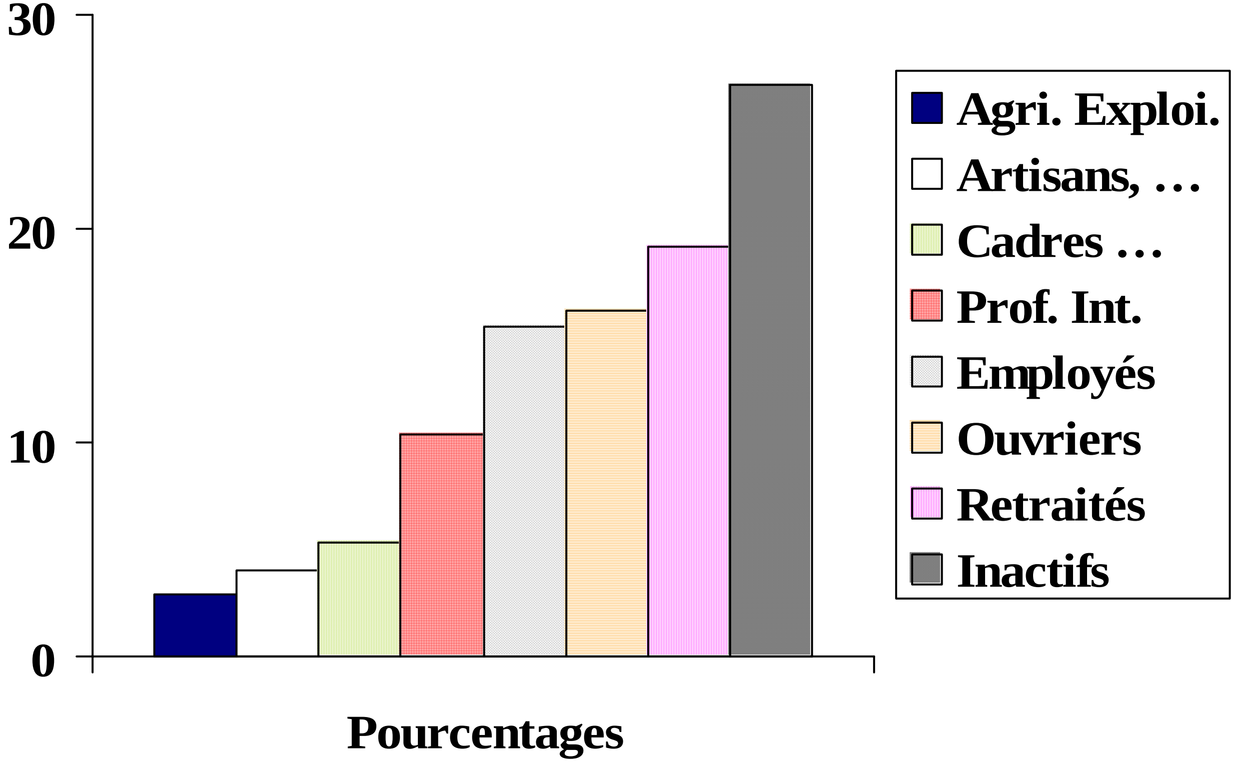
Règle: sur les graphiques, les aires des modalités sont proportionnelles à leurs effectifs

2.4 Représentations graphiques:

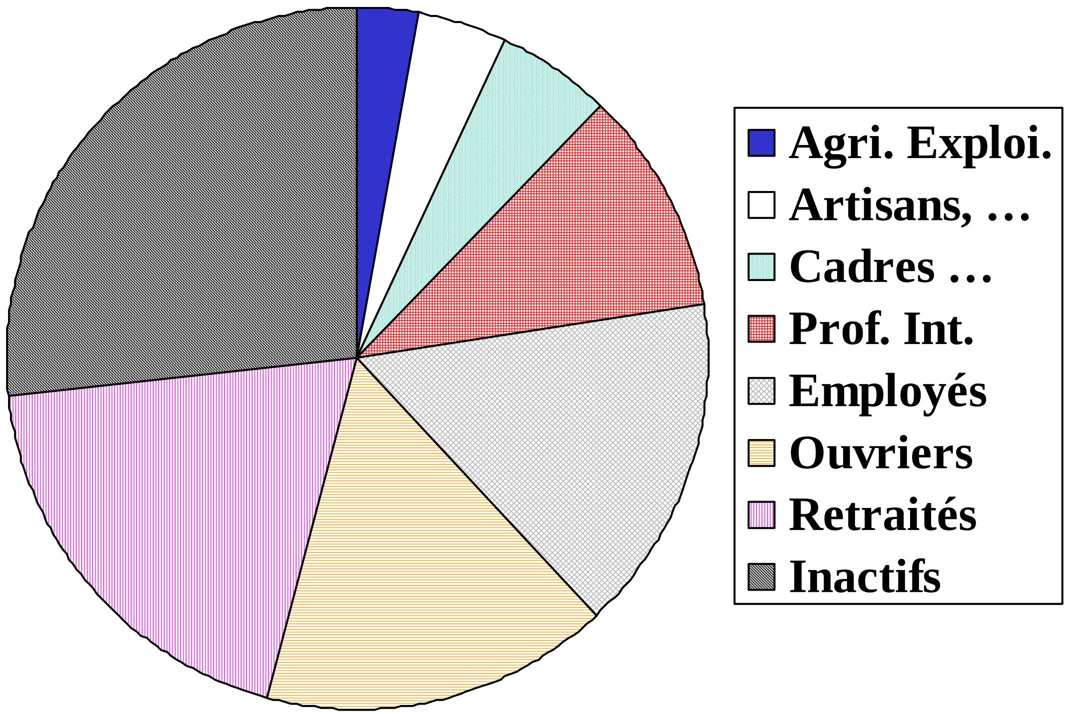
Règle: sur les graphiques, les aires des modalités sont proportionnelles à leurs effectifs

a. Diagramme en barre:

La hauteur des barres est proportionnelle à l'effectif de la modalité



b. Diagramme en secteurs: L'angle du secteur de disque est proportionnel à l'effectif de la modalité



3. Étude d'une variable quantitative discrète

Ménage Français par rapport à leur effectif en 1989

Nbe personnes	Effectif	Pourcentage
1 personne	7079434	31.6
2 personnes	7086664	31.6
3 personnes	3619655	16.1
4 personnes	3057674	13.6
5 personnes	1182235	5.3
6 ou plus	109189	1.8
Total	22434621	100

3. Étude d'une variable quantitative discrète

Ménage Français par rapport à leur effectif en 1989

Nbe personnes	Effectif	Pourcentage
1 personne	7079434	31.6
2 personnes	7086664	31.6
3 personnes	3619655	16.1
4 personnes	3057674	13.6
5 personnes	1182235	5.3
6 ou plus	109189	1.8
Total	22434621	100

On considère 6 et + comme valant 6

3.1 Fréquence cumulée: proportion d'individus dont la valeur du caractère est inférieure ou égale à la valeur considérée

Nbe pers.	Effectif	*Pi*	F. Cumulée en %
1 pers.	7079434	32	32
2 pers.	7086664	32	
3 pers.	3619655	16	
4 pers.	3057674	14	
5 pers.	1182235	5	
6 ou plus	109189	2	
Total	22434621	100	

3.1 Fréquence cumulée: proportion d'individus dont la valeur du caractère est inférieure ou égale à la valeur considérée

Nbe pers.	Effectif	*Pi*	F. Cumulée en %
1 pers.	7079434	32	32
2 pers.	7086664	32	63
3 pers.	3619655	16	
4 pers.	3057674	14	
5 pers.	1182235	5	
6 ou plus	109189	2	
Total	22434621	100	

$$\frac{7079434 + 7086664}{22434621}$$

$$32+32=64$$

3.1 Fréquence cumulée: proportion d'individus dont la valeur du caractère est inférieure ou égale à la valeur considérée

Nbe pers.	Effectif	*Pi*	F. Cumulée en %
1 pers.	7079434	32	32
2 pers.	7086664	32	63
3 pers.	3619655	16	79
4 pers.	3057674	14	93
5 pers.	1182235	5	98
6 ou plus	109189	2	100
Total	22434621	100	

$$\frac{7079434 + 7086664}{22434621}$$

$$32+32=64$$

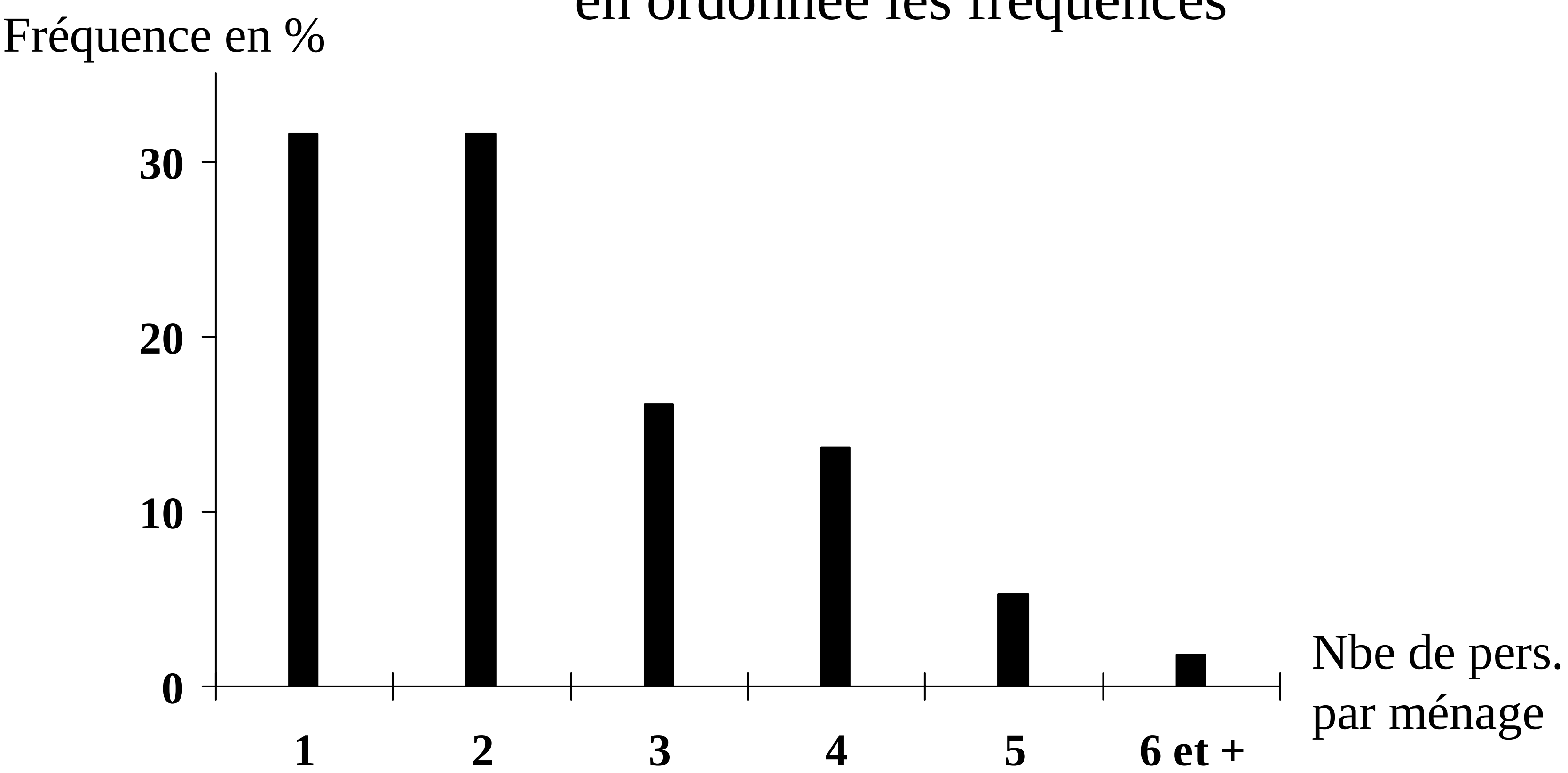
En 1989, 63% des ménages sont composés de 2 personnes ou moins

3.2 Représentations graphiques:

a. Histogramme des fréquences:

Diagramme en bâton: en abscisse les valeurs du caractère

en ordonnée les fréquences

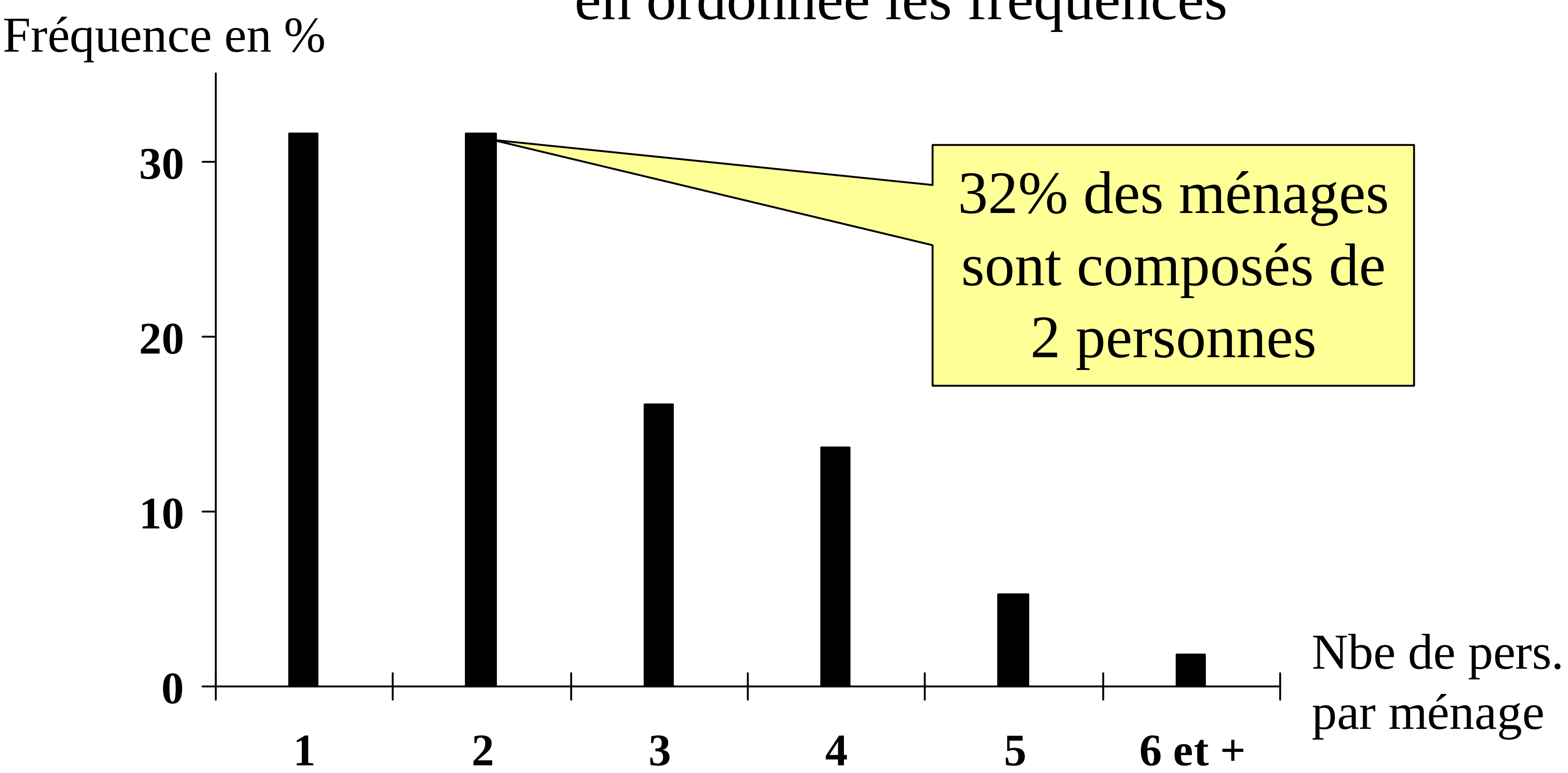


3.2 Représentations graphiques:

a. Histogramme des fréquences:

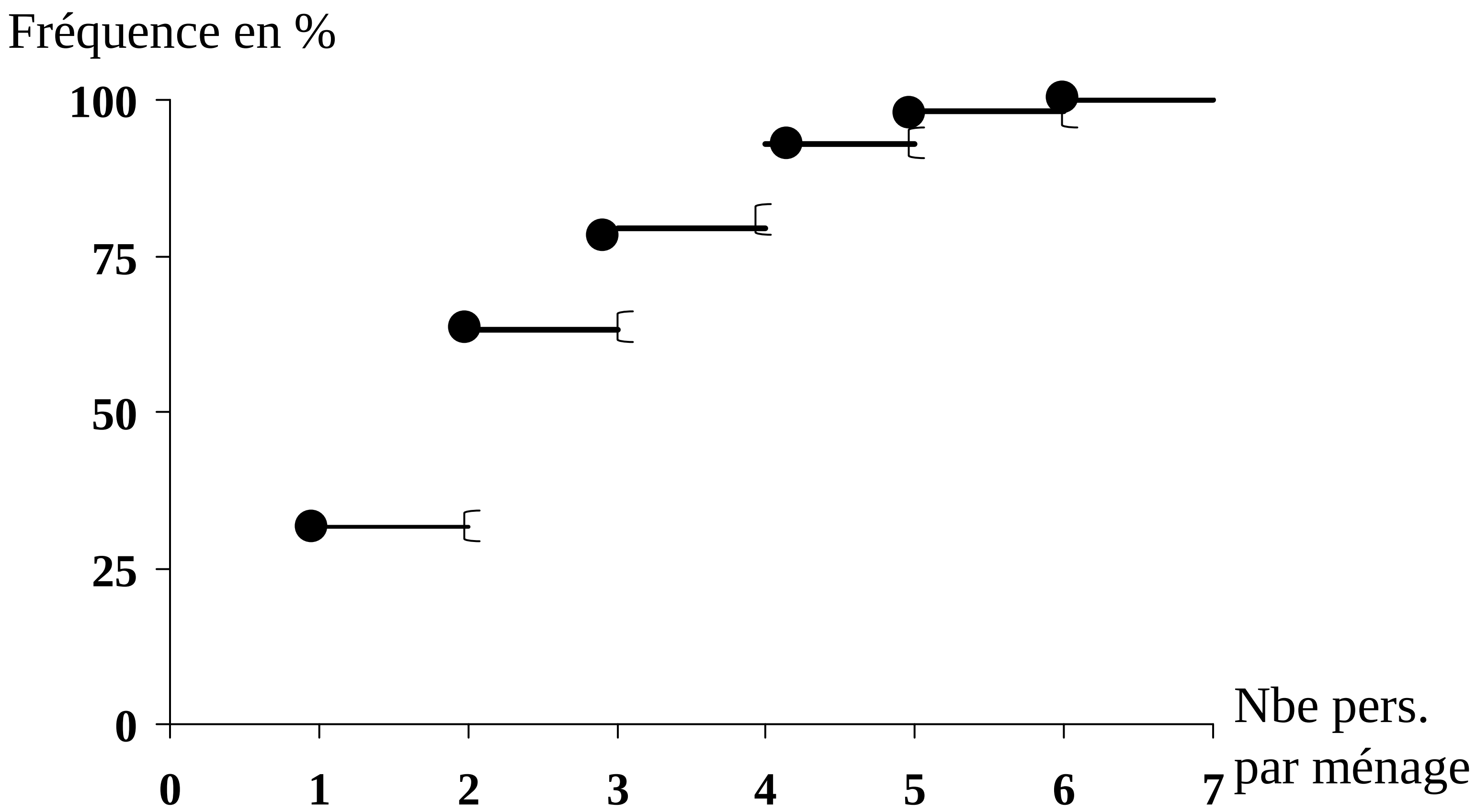
Diagramme en bâton: en abscisse les valeurs du caractère

en ordonnée les fréquences



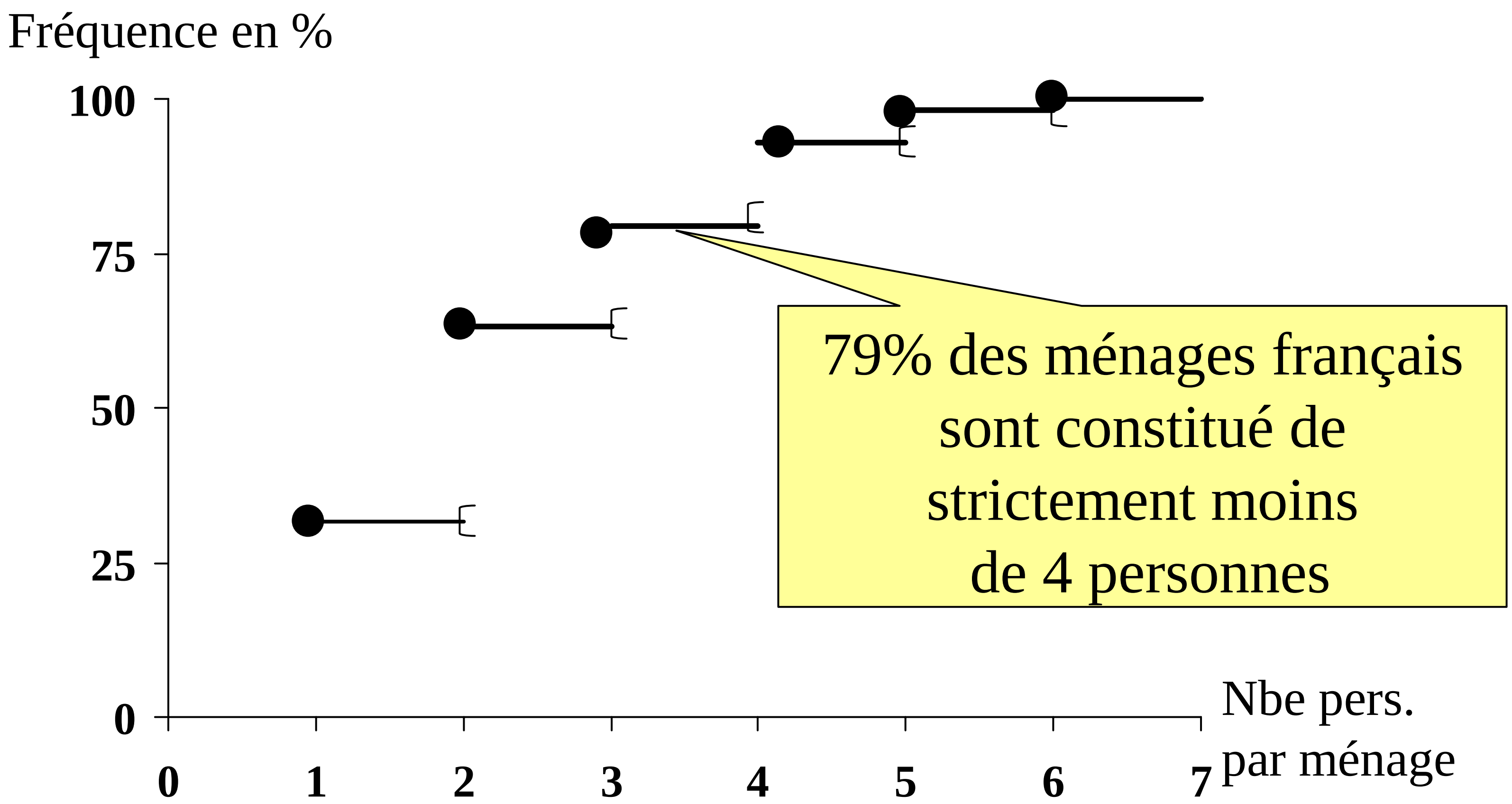
b. Diagramme cumulatif:

Représente les fréquences cumulées en fonction des valeurs du caractère



b. Diagramme cumulatif:

Représente les fréquences cumulées en fonction des valeurs du caractère



3.3 Résumé numérique d'une distribution:

a. Caractéristiques centrales:

La moyenne notée $\overline{x}$

Moyenne arithmétique des valeurs du caractère pour les *n* individus de la population

$$\overline{x} = \frac{1}{n} \sum_i n_i x_i = \sum_i f_i x_i$$

3.3 Résumé numérique d'une distribution:

a. Caractéristiques centrales:

La moyenne notée $\overline{x}$

Représente le **barycentre** des valeurs prises par le caractère

Moyenne arithmétique des valeurs du caractère pour les *n* individus de la population

$$\overline{x} = \frac{1}{n}\sum_i n_i x_i = \sum_i f_i x_i$$

DOYEN

$$\bar{x} = \frac{1}{n}\sum_i n_i\, x_i = \sum_i f_i\, x_i$$

Nbe pers.	Effectif	*Pi*
1 pers.	7079434	32
2 pers.	7086664	32
3 pers.	3619655	16
4 pers.	3057674	14
5 pers.	1182235	5
6 ou plus	109189	2
Total	22434621	100

$\bar{x} =$

0.32*1

+0.32*2

+0.16*3

+0.14*4

+0.05*5

+0.02*6

≈ 2.4 (personnes)

DOYEN

$$\bar{x} = \frac{1}{n}\sum_i n_i x_i = \sum_i f_i x_i$$

Nbe pers.	Effectif	*Pi*
1 pers.	7079434	32
2 pers.	7086664	32
3 pers.	3619655	16
4 pers.	3057674	14
5 pers.	1182235	5
6 ou plus	109189	2
Total	22434621	100

$\bar{x} =$

0.32*1

+0.32*2

+0.16*3

+0.14*4

+0.05*5

+0.02*6

≈2.4 (personnes)

Ne pas oublier l'unité

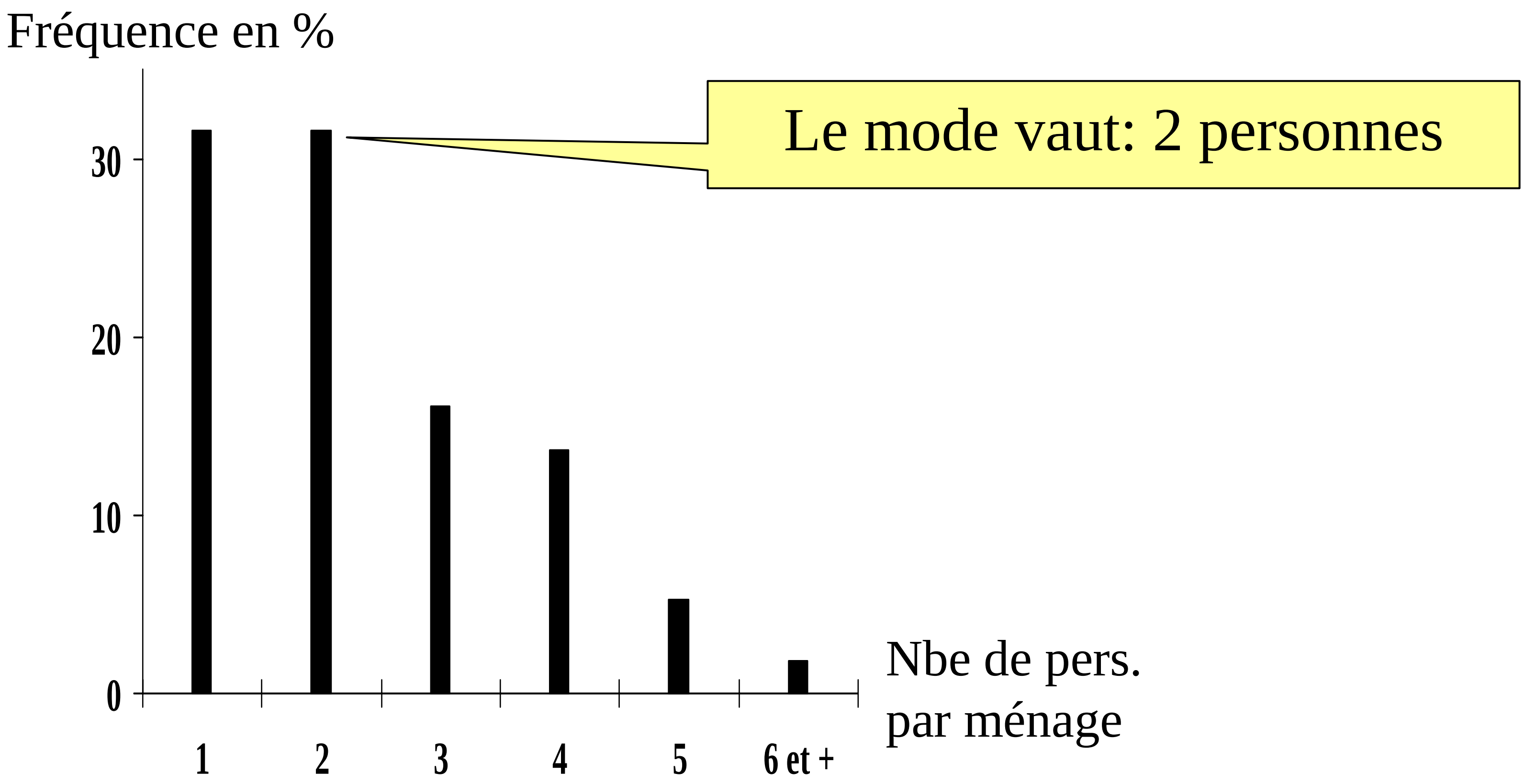
En 1989 en France, il y a en moyenne 2.4 personnes par ménage

Le(s) mode(s)

Valeurs du caractère en lesquelles l’histogramme des fréquences possède un maximum relatif

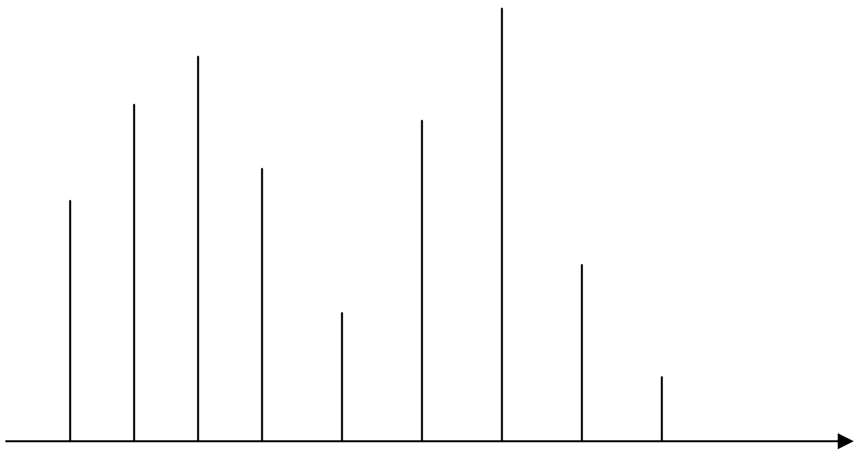
Le(s) mode(s)

Valeurs du caractère en lesquelles l'histogramme des fréquences possède un maximum relatif



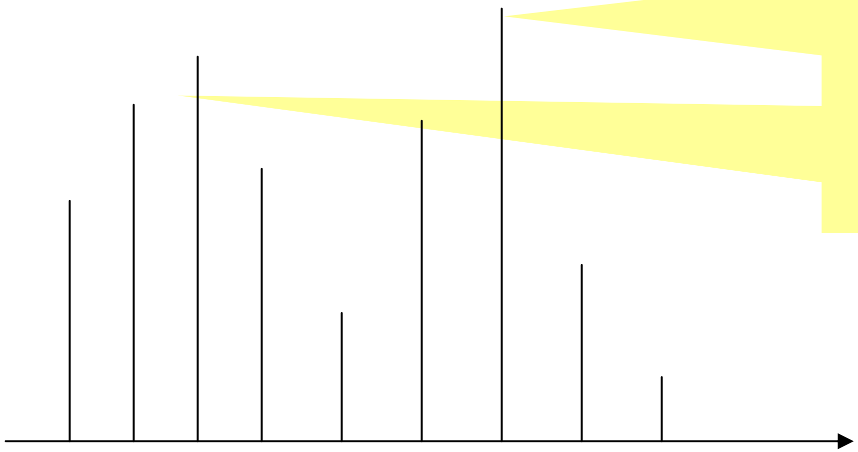
Le mode

Valeurs du caractère en lesquels l'histogramme des fréquences possède un maximum **RELATIF**



Le mode

Valeurs du caractère en lesquels l'histogramme des fréquences possède un maximum **RELATIF**



Cette distribution a
2 modes!
Elle est BIMODALE

C'est souvent caractéristique d'une population
NON HOMOGENE

La médiane

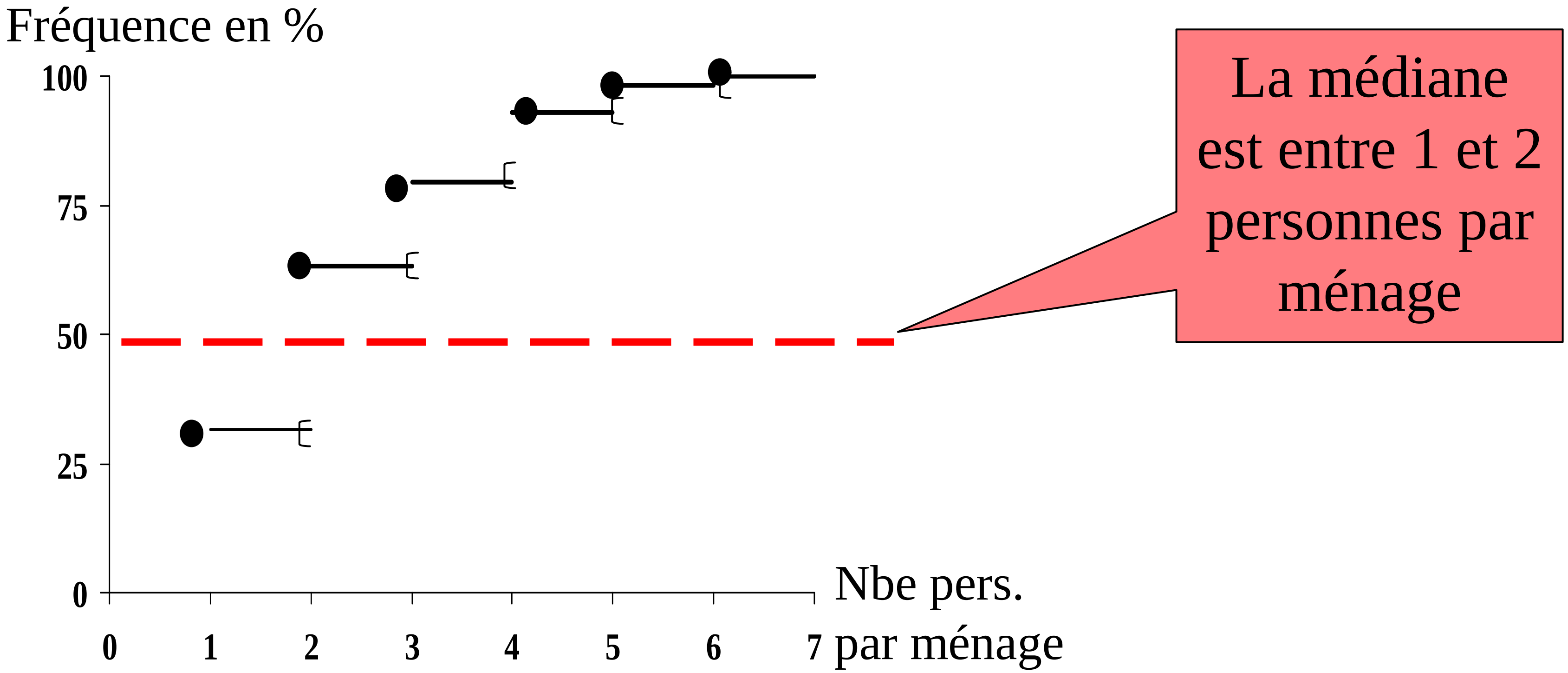
Valeur du caractère qui partage la série statistique en 2 groupes de même fréquence (0.5).

On la détermine à l'aide des fréquences cumulées ou du diagramme cumulatif

La médiane

Valeur du caractère qui partage la série statistique en 2 groupes de même fréquence (0.5).

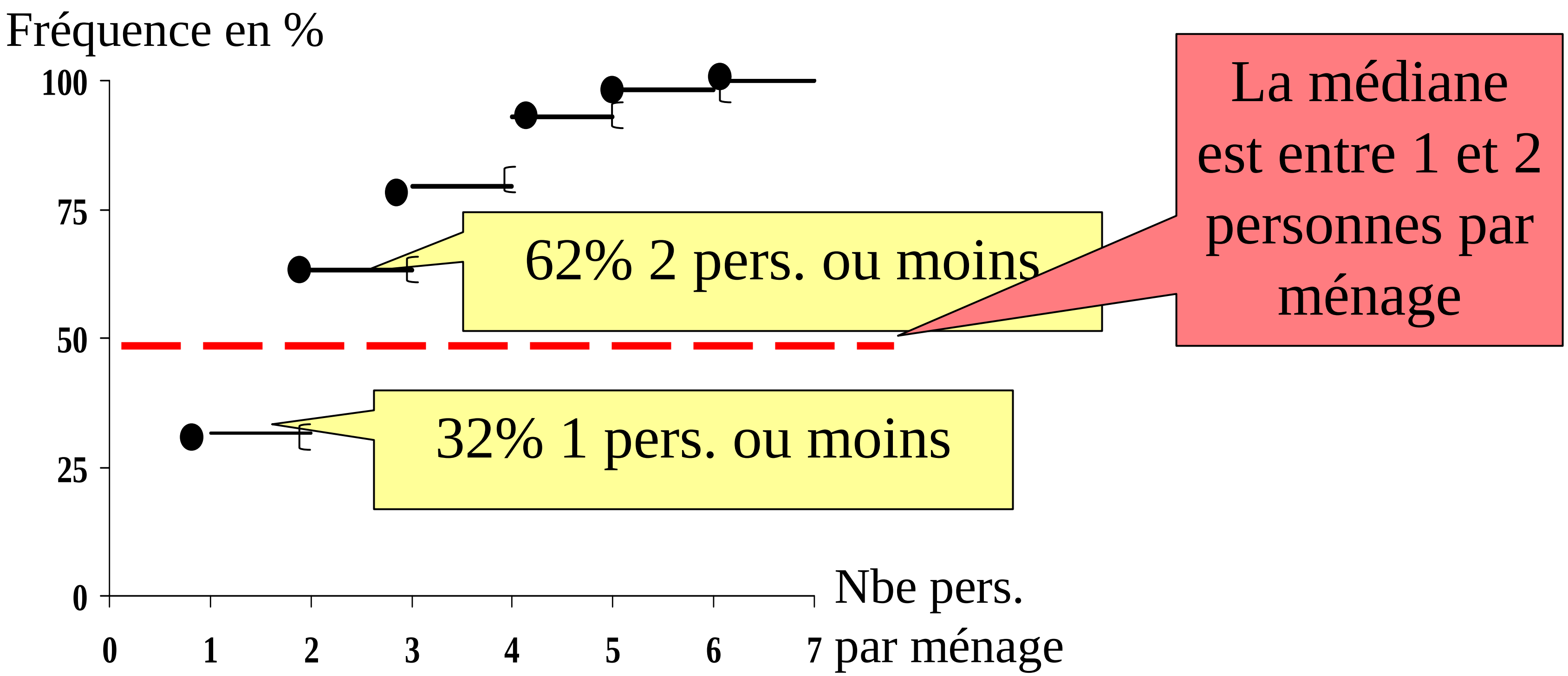
On la détermine à l'aide des fréquences cumulées ou du diagramme cumulatif



La médiane

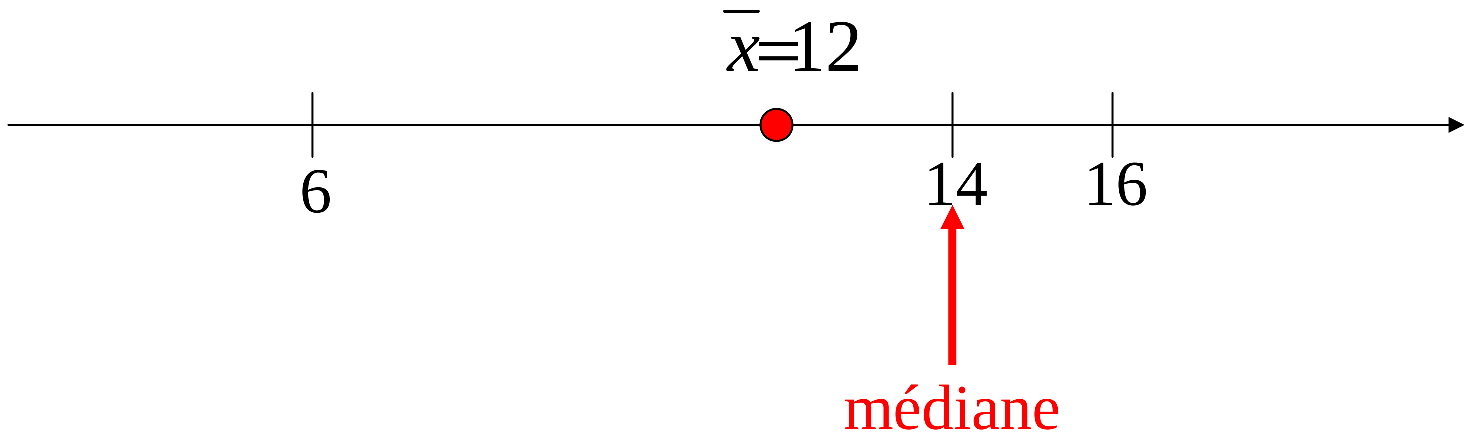
Valeur du caractère qui partage la série statistique en 2 groupes de même fréquence (0.5).

On la détermine à l'aide des fréquences cumulées ou du diagramme cumulatif



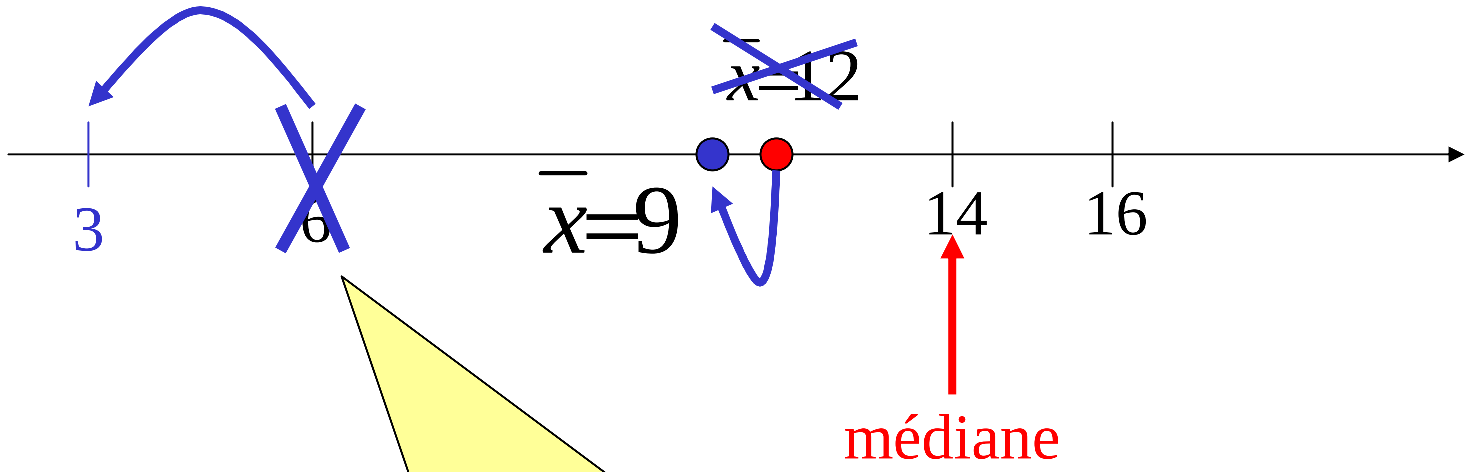
Quelle est la différence entre moyenne et médiane?

Note de préparation à la maison semaine3:



Quelle est la différence entre moyenne et médiane?

Note de préparation à la maison semaine3:



La médiane est peu sensible aux valeurs aberrantes contrairement à la moyenne

b. Caractéristiques de dispersion:

Exemple:

Notes des devoirs à la maison en 2001 à l'IUP com et vente

- Semaine 1: 9, 10, 10, 11
- Semaine 2: 0, 10, 10, 20

b. Caractéristiques de dispersion:

Exemple:

Notes des devoirs à la maison en 2001 à l'IUP com et vente

- Semaine 1: 9, 10, 10, 11
- Semaine 2: 0, 10, 10, 20

Toutes les caractéristiques centrales valent 10!

b. Caractéristiques de dispersion:

Exemple:

Notes des devoirs à la maison en 2001 à l'IUP com et vente

- Semaine 1: 9, 10, 10, 11
- Semaine 2: 0, 10, 10, 20

Toutes les caractéristiques centrales valent 10!

Trouver des valeurs numériques qui caractérisent la dispersion de la distribution

➡ Comment les valeurs sont elles éloignées de la moyenne?

Une mauvaise idée: $$\frac{1}{n}\sum_i n_i(x_i - \bar{x})$$

- Semaine 1: 9, 10, 10, 11

$$\frac{1}{4}\left(1*(9-10) + 2*(10-10) + 1*(11-10)\right) = 0$$

<u>Une mauvaise idée</u>: $\frac{1}{n}\sum_i n_i(x_i - \bar{x})$

•Semaine 1: 9, 10, 10, 11

$$\frac{1}{4}\left(1*(9-10) + 2*(10-10) + 1*(11-10) \right) = 0$$

$$+$$
$$||$$
$$0$$

Les écarts positifs et négatifs se compensent!

L'écart absolu moyen:

La moyenne des ECARTS ABSOLUS à la moyenne

$$\overline{e}_{\overline{x}} = \frac{1}{n}\sum_i n_i \left| x_i - \overline{x} \right| = \sum_i f_i \left| x_i - \overline{x} \right|$$

L'écart absolu moyen:

$$\bar{x} = 2.4 \text{ (personnes)}$$

Nb pers.	Effectif	*Pi*
1 pers.	7079434	32
2 pers.	7086664	32
3 pers.	3619655	16
4 pers.	3057674	14
5 pers.	1182235	5
6 ou plus	109189	2
Total	22434621	100

$$0.32 * |1-2.4|$$
$$+ 0.32 * |2-2.4|$$
$$+ 0.16 * |3-2.4|$$
$$+ 0.14 * |4-2.4|$$
$$+ 0.05 * |5-2.4|$$
$$+ 0.02 * |6-2.4|$$

$$\bar{e}_{\bar{x}} \approx 1.4 \text{ (personnes)}$$

L'écart absolu moyen:

$$\bar{x} = 2.4 \text{ (personnes)}$$

Nb pers.	Effectif	*Pi*
1 pers.	7079434	32
2 pers.	7086664	32
3 pers.	3619655	16
4 pers.	3057674	14
5 pers.	1182235	5
6 ou plus	109189	2
Total	22434621	100

$$\begin{aligned} & 0.32 * |1-2.4| \\ +\ & 0.32 * |2-2.4| \\ +\ & 0.16 * |3-2.4| \\ +\ & 0.14 * |4-2.4| \\ +\ & 0.05 * |5-2.4| \\ +\ & 0.02 * |6-2.4| \end{aligned}$$

$$\bar{e}_{\bar{x}} \approx 1.4 \text{ (personnes)}$$

Attention à l'unité

La variance et l'écart-type:

La variance est la moyenne des carrés des écarts à la moyenne

$$\sigma^2 = \frac{1}{n}\sum_i n_i \left(x_i - \overline{x}\right)^2 = \sum_i f_i \left(x_i - \overline{x}\right)^2$$

La variance et l'écart-type:

La variance est la moyenne des carrés des écarts à la moyenne

Si x a pour unité la personne, alors σ^2 a pour unité personne^2

$$\sigma^2 = \frac{1}{n}\sum_i n_i \left(x_i - \overline{x}\right)^2 = \sum_i f_i \left(x_i - \overline{x}\right)^2$$

La variance et l'écart-type:

La variance est la moyenne des carrés des écarts à la moyenne

Si x a pour unité la personne, alors σ^2 a pour unité personne^2

$$\sigma^2 = \frac{1}{n}\sum_i n_i \left(x_i - \overline{x}\right)^2 = \sum_i f_i \left(x_i - \overline{x}\right)^2$$

L'écart-type est la racine carré de la variance

$$\sigma = \sqrt{\sigma^2}$$

La variance et l'écart-type:

La variance est la moyenne des carrés des écarts à la moyenne

Si x a pour unité la personne, alors σ^2 a pour unité personne^2

$$\sigma^2 = \frac{1}{n}\sum_i n_i\left(x_i - \overline{x}\right)^2 = \sum_i f_i\left(x_i - \overline{x}\right)^2$$

L'écart-type est la racine carré de la variance

Même unité que le caractère

$$\sigma = \sqrt{\sigma^2}$$

La variance et l'écart-type:

La variance est la moyenne des carrés des écarts à la moyenne

Si x a pour unité la personne, alors σ^2 a pour unité personne^2

$$\sigma^2 = \frac{1}{n}\sum_i n_i \left(x_i - \overline{x}\right)^2 = \sum_i f_i \left(x_i - \overline{x}\right)^2$$

L'écart-type est la racine carré de la variance

Même unité que le caractère

$$\sigma = \sqrt{\sigma^2}$$

Entre $\overline{x} - 2\sigma$ et $\overline{x} + 2\sigma$ il y a au moins 75% de la population

Pour calculer la variance on peut utiliser la formule:

$$\sigma^2 = \left(\sum_i f_i \, x_i^2 \right) - \overline{x}^2$$

Nbe pers.	Effectif	*Pi*
1 pers.	7079434	32
2 pers.	7086664	32
3 pers.	3619655	16
4 pers.	3057674	14
5 pers.	1182235	5
6 ou plus	109189	2
Total	22434621	100

$\overline{x} = 2.4$ (personnes)

$0.32 * 1^2$
$+ 0.32 * 2^2$
$+ 0.16 * 3^2$
$+ 0.14 * 4^2$
$+ 0.05 * 5^2$
$+ 0.02 * 6^2$

$\sigma^2 \approx 7.25 - 2.4^2 \approx 1.5$ (personnes^2)

Pour calculer la variance on peut utiliser la formule:

$$\sigma^2 = \left(\sum_i f_i x_i^2 \right) - \bar{x}^2$$

Nbe pers.	Effectif	*Pi*
1 pers.	7079434	32
2 pers.	7086664	32
3 pers.	3619655	16
4 pers.	3057674	14
5 pers.	1182235	5
6 ou plus	109189	2
Total	22434621	100

$\bar{x} = 2.4$ (personnes)

$0.32 * 1^2$

$+ 0.32 * 2^2$

$+ 0.16 * 3^2$

$+ 0.14 * 4^2$

$+ 0.05 * 5^2$

$+ 0.02 * 6^2$

Attention à l'unité

$\sigma^2 \approx 7.25 - 2.4^2 \approx 1.5$ (personnes^2)

$$\sigma \approx \sqrt{1.5} \approx 1.2 \text{ (personne)}$$

En 1999, au moins 75% des ménages français ont un effectif entre 0 et 4.8 personnes.

4. Étude d'une variable quantitative continue

Même notion que dans le chapitre précédent.
La seule différence est que on ne considère pas les modalité une par une mais par CLASSES

4. Étude d'une variable quantitative continue

Même notion que dans le chapitre précédent.
La seule différence est que on ne considère pas les modalité une par une mais par CLASSES

Intervalle de valeurs possibles pour la variable statistique continue

Population française active par âge en 1999

Age	Effectif	Pourcentage	Cumul
15-24	2279542	8.6	8.6
25-29	3628502	13.7	22.3
30-34	3771554	14.2	36.5
35-39	3865252	14.6	51.0
40-44	3770300	14.2	65.2
45-49	3696642	13.9	79.2
50-54	3305278	12.5	91.6
55 et +	2225411	8.4	100
Total	26542481	100	100

Population française active par âge en 1999

Age	Effectif	Pourcentage	Cumul
15-24	2279542	8.6	8.6
25-29	3628502	13.7	22.3
30-34	3771554	14.2	36.5
35-39	3865252	14.6	51.0
40-44	3770300	14.2	65.2
45-49	3696642	13.9	79.2
50-54	3305278	12.5	91.6
55 et +	2225411	8.4	100
Total	26542481	100	100

Il y a 3771554 personnes dans la classe d'âge des 30-34 ans

Comment déterminer les classes?

Comment déterminer les classes?

- Nombre de classes relativement faible: ≤ 10

Comment déterminer les classes?

- Nombre de classes relativement faible: ≤ 10

- Effectif des classes du même ordre de grandeur

  Classe fine là où le caractère est plus fréquent

  Classe large là où le caractère est rare

Comment déterminer les classes?

- Nombre de classes relativement faible: ≤ 10

- Effectif des classes du même ordre de grandeur

  Classe fine là où le caractère est plus fréquent

  Classe large là où le caractère est rare

- Essayer d'utiliser des classes de même amplitude

Comment déterminer les classes?

- Nombre de classes relativement faible: ≤ 10

- Effectif des classes du même ordre de grandeur

  Classe fine là où le caractère est plus fréquent

  Classe large là où le caractère est rare

- Essayer d'utiliser des classes de même amplitude

Souvent la première et la dernière classe n'ont pas la même amplitude

4.1 Fréquence relative

Quand les amplitudes des classes sont différentes on ne considère plus les fréquences, mais les FREQUENCES RELATIVES:

$$\frac{f_i}{a_i}$$

a_i est l'amplitude de la classe

.ai	Age	Effectif	*.fi*	Cumul	*.f relative à 5 ans*
2	15-24	2279542	0.086	8.6	0.043
1	25-29	3628502	0.137	22.3	0.137
1	30-34	3771554	0.142	36.5	0.142
1	35-39	3865252	0.146	51.0	0.146
1	40-44	3770300	0.142	65.2	0.142
1	45-49	3696642	0.139	79.2	0.139
1	50-54	3305278	0.125	91.6	0.125
2	55 et +	2225411	0.084	100	0.042
	Total	26542481	1	100	

.ai	Age	Effectif	*.fi*	Cumul	*.f relative à 5 ans*
2	15-24	2279542	0.086	8.6	0.043
1	25-29	3628502	0.137	22.3	0.137
1	30-34	3771554	0.142	36.5	0.142
1	35-39	3865252	0.146	51.0	0.146
1	40-44	3770300	0.142	65.2	0.142
1	45-49	3696642	0.139	79.2	0.139
1	50-54	3305278	0.125	91.6	0.125
2	55 et +	2225411	0.084	100	0.042
	Total	26542481	1	100	

Pour avoir la largeur de classe il faut fixer la borne supérieur de la classe.
Il faut prendre une décision raisonnable.
Ici on parle de population active: 55-64

4.2 Représentations graphiques:

a. Histogramme des fréquences:

Les classes de la distribution forment les bases des batons

Les **SURFACES** sont proportionnelles aux fréquences!

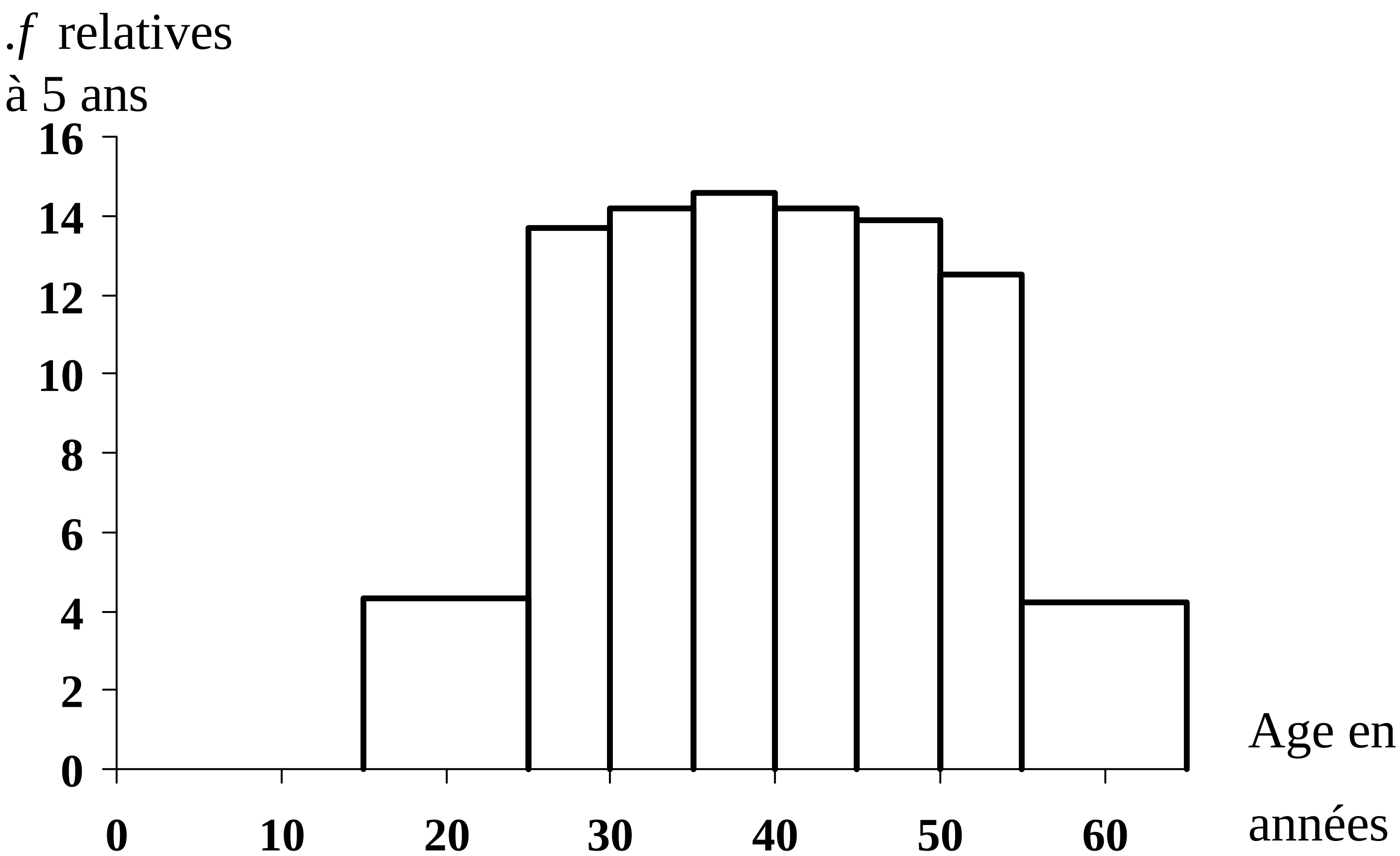
4.2 Représentations graphiques:

a. Histogramme des fréquences:

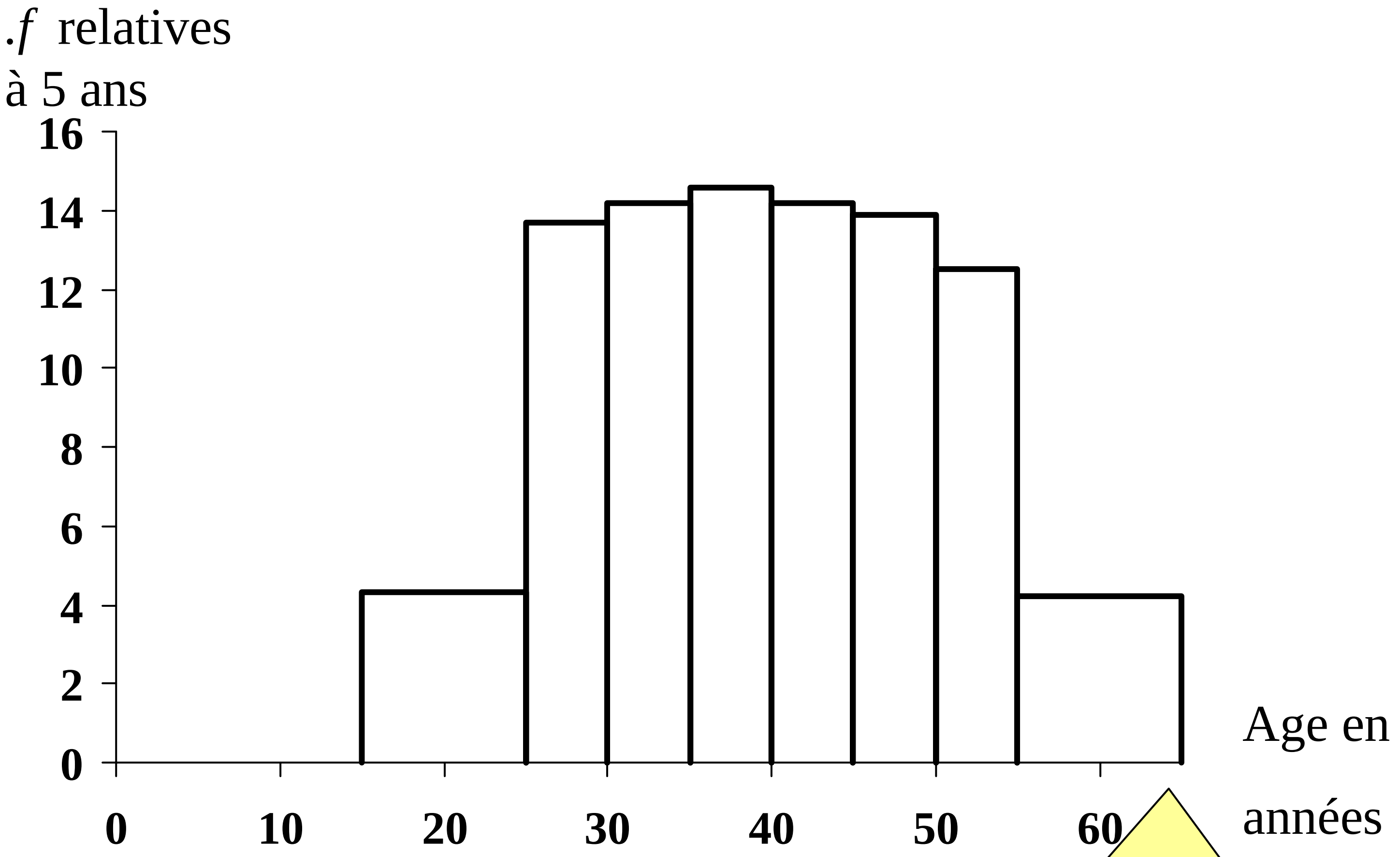
Les classes de la distribution forment les bases des batons

Les **SURFACES** sont proportionnelles aux fréquences!

Donc si les classes sont d'amplitudes différentes, les **HAUTEURS** des histogrammes sont proportionnelles aux **FREQUENCES RELATIVES**.



.*f* relatives
à 5 ans
16
14
12
10
8
6
4
2
0
0
10
20
30
40
50
60
Age en
années



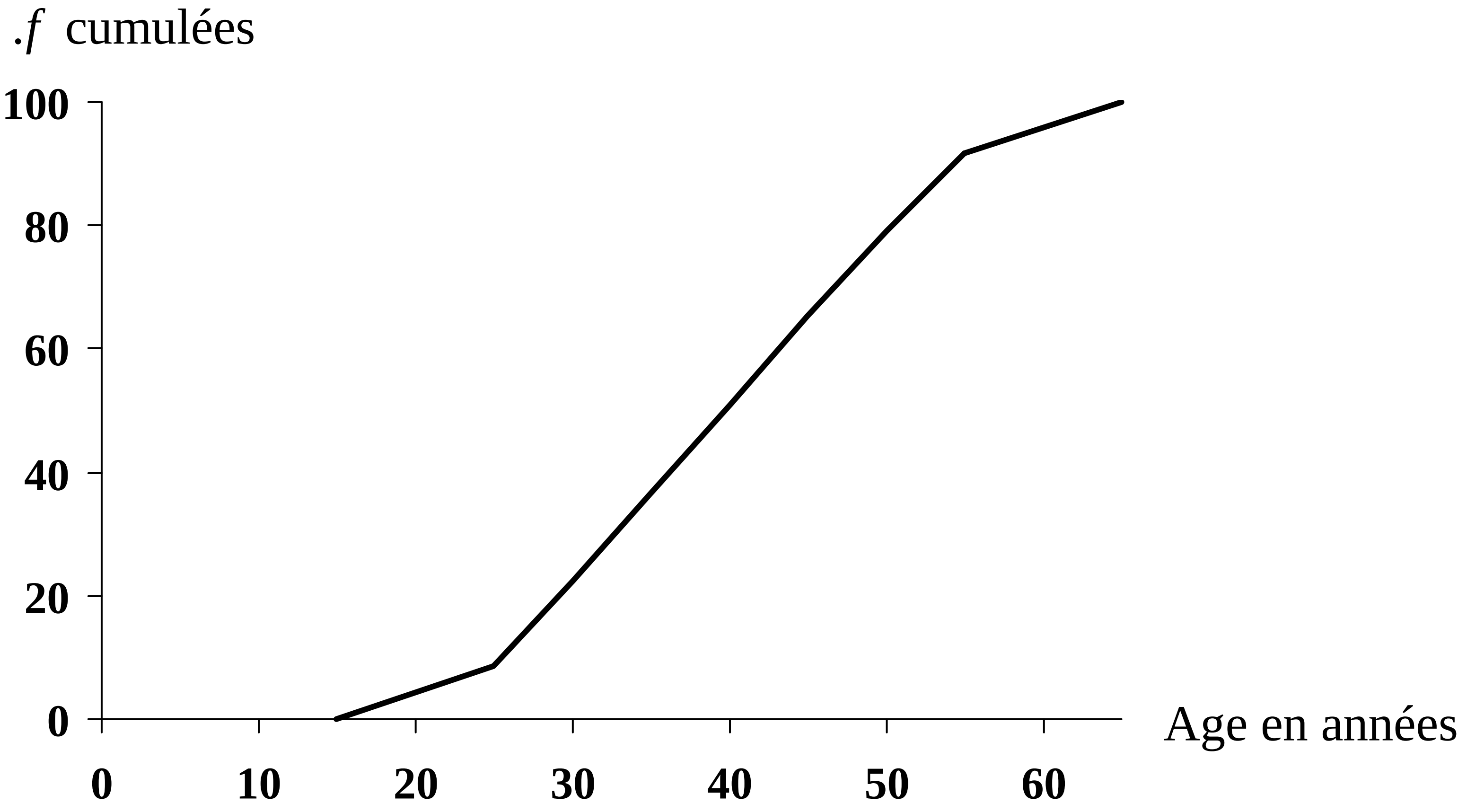
Pour la borne supérieure on conserve toujours la même

b. Polygone des fréquences cumulées:

En abscisse les limites de classes

En ordonnée les fréquence cumulées

On rejoint les points par une ligne brisée



4.3 Résumé numérique d'une distribution:

a. Caractéristiques centrales:

La moyenne notée $\overline{x}$

Moyenne arithmétique des valeurs du caractère pour les *n* individus de la population

$$\overline{x} = \frac{1}{n}\sum_i n_i c_i = \sum_i f_i c_i$$

4.3 Résumé numérique d'une distribution:

a. Caractéristiques centrales:

La moyenne notée $\overline{x}$

Moyenne arithmétique des valeurs du caractère pour les *n* individus de la population

$$\overline{x} = \frac{1}{n}\sum_i n_i c_i = \sum_i f_i c_i$$

On ne considère plus les valeurs des modalités, mais les **CENTRES DES CLASSES**

4.3 Résumé numérique d'une distribution:

a. Caractéristiques centrales:

La moyenne notée $\overline{x}$

Représente le **barycentre** des valeurs prises par le caractère

Moyenne arithmétique des valeurs du caractère pour les *n* individus de la population

$$\overline{x} = \frac{1}{n}\sum_i n_i c_i = \sum_i f_i c_i$$

On ne considère plus les valeurs des modalités, mais les **CENTRES DES CLASSES**

DOYEN

$$\bar{x} = \sum_i f_i c_i$$

Age	Effectif	*.fi*	Cumul
15-24	2279542	0.086	8.6
25-29	3628502	0.137	22.3
30-34	3771554	0.142	36.5
35-39	3865252	0.146	51.0
40-44	3770300	0.142	65.2
45-49	3696642	0.139	79.2
50-54	3305278	0.125	91.6
55 et +	2225411	0.084	100
Total	26542481	1	100

DOYEN

$$\bar{x} = \sum_i f_i c_i$$

$$\frac{15+24}{2} \approx 20$$

$.ci$	Age	Effectif	$.fi$	Cumul
20	15-24	2279542	0.086	8.6
27	25-29	3628502	0.137	22.3
33	30-34	3771554	0.142	36.5
37	35-39	3865252	0.146	51.0
43	40-44	3770300	0.142	65.2
47	45-49	3696642	0.139	79.2
53	50-54	3305278	0.125	91.6
60	55 et +	2225411	0.084	100
	Total	26542481	1	100

DOYEN

$$\overline{x} = \sum_i f_i c_i$$

$$\frac{15+24}{2} \approx 20$$

.ci	Age	Effectif	.fi	Cumul
20	15-24	2279542	0.086	8.6
27	25-29	3628502	0.137	22.3
33	30-34	3771554	0.142	36.5
37	35-39	3865252	0.146	51.0
43	40-44	3770300	0.142	65.2
47	45-49	3696642	0.139	79.2
53	50-54	3305278	0.125	91.6
60	55 et +	2225411	0.084	100
	Total	26542481	1	100

0.086*20
+ 0.137*27
+ 0.142*33
+ 0.146*37
+ 0.142*43
+ 0.139*47
+ 0.125*53
+ 0.042*60

$$\overline{x} \approx 40 \text{ (ans)}$$

$$\bar{x} = \sum_i f_i c_i$$

$$\frac{15+24}{2} \approx 20$$

.ci	Age	Effectif	.fi	Cumul
20	15-24	2279542	0.086	8.6
27	25-29	3628502	0.137	22.3
33	30-34	3771554	0.142	36.5
37	35-39	3865252	0.146	51.0
43	40-44	3770300	0.142	65.2
47	45-49	3696642	0.139	79.2
53	50-54	3305278	0.125	91.6
60	55 et +	2225411	0.084	100
	Total	26542481	1	100

0.086*20
+ 0.137*27
+ 0.142*33
+ 0.146*37
+ 0.142*43
+ 0.139*47
+ 0.125*53
+ 0.042*60

Ne pas oublier l'unité

$$\bar{x} \approx 40 \text{ (ans)}$$

En 1999 en France, les actifs ont une moyenne d'âge de 40 ans

Classe(s) modale(s)

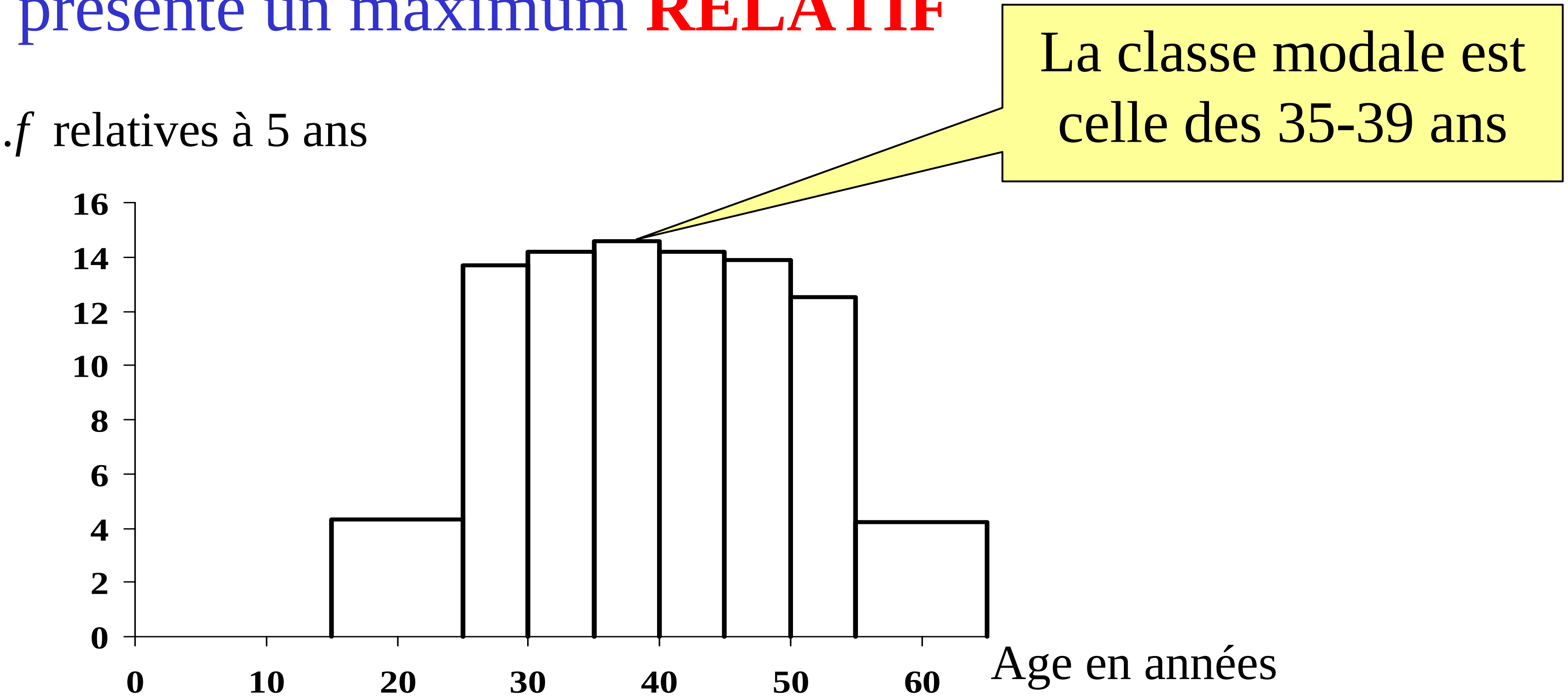
CLASSES en lesquelles l'histogramme des fréquences présente un maximum RELATIF

Classes en laquelle la fréquence **RELATIVE** présente un maximum **RELATIF**

Classe(s) modale(s)

CLASSES en lesquelles l'histogramme des fréquences présente un maximum RELATIF

Classes en laquelle la fréquence **RELATIVE** présente un maximum **RELATIF**



La médiane

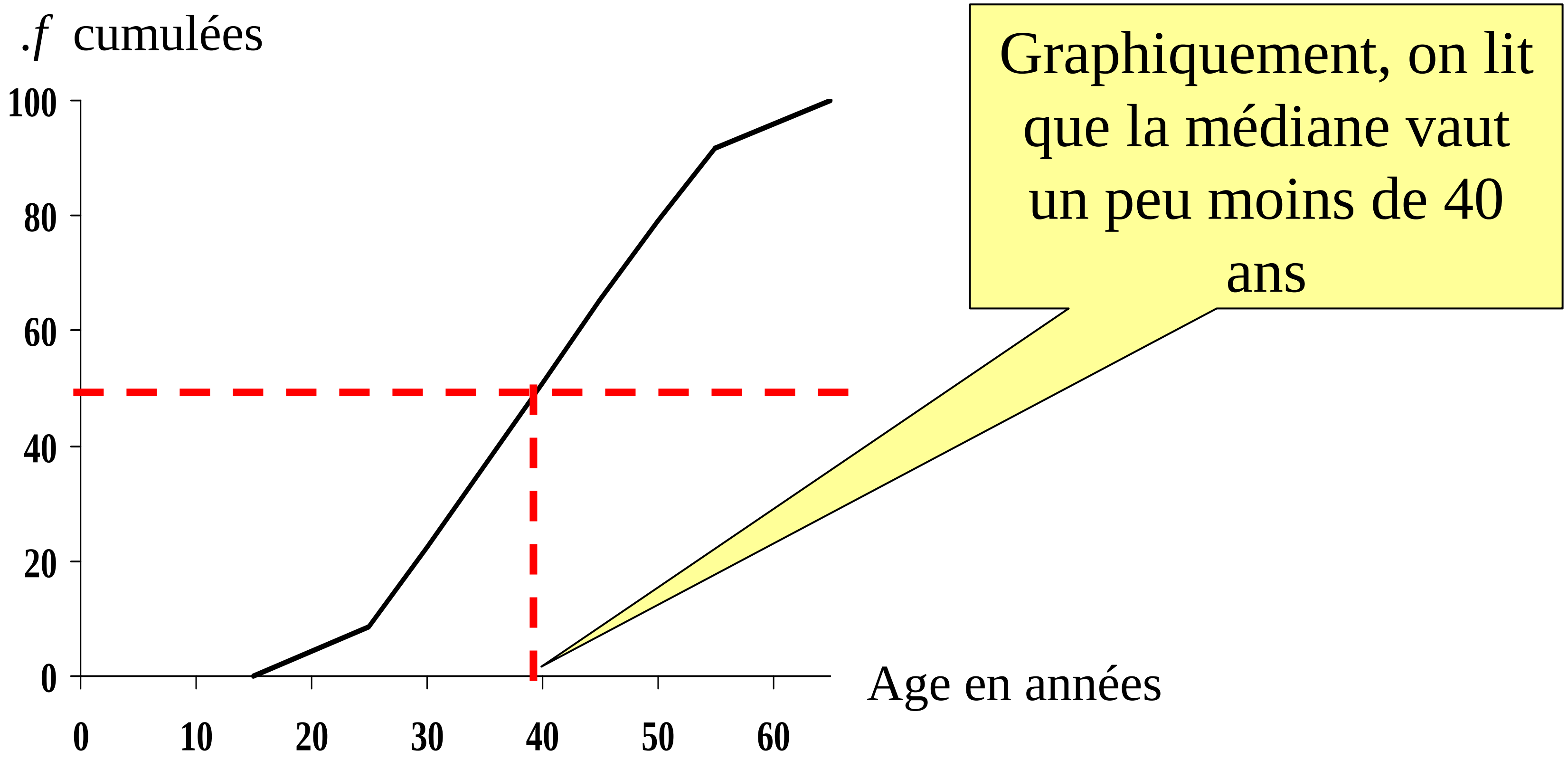
Valeur du caractère qui partage la série statistique en 2 groupes de même fréquence (0.5).

C'est la valeur correspondant à un effectif cumulé de 50% sur le polygone des fréquences cumulées

La médiane

Valeur du caractère qui partage la série statistique en 2 groupes de même fréquence (0.5).

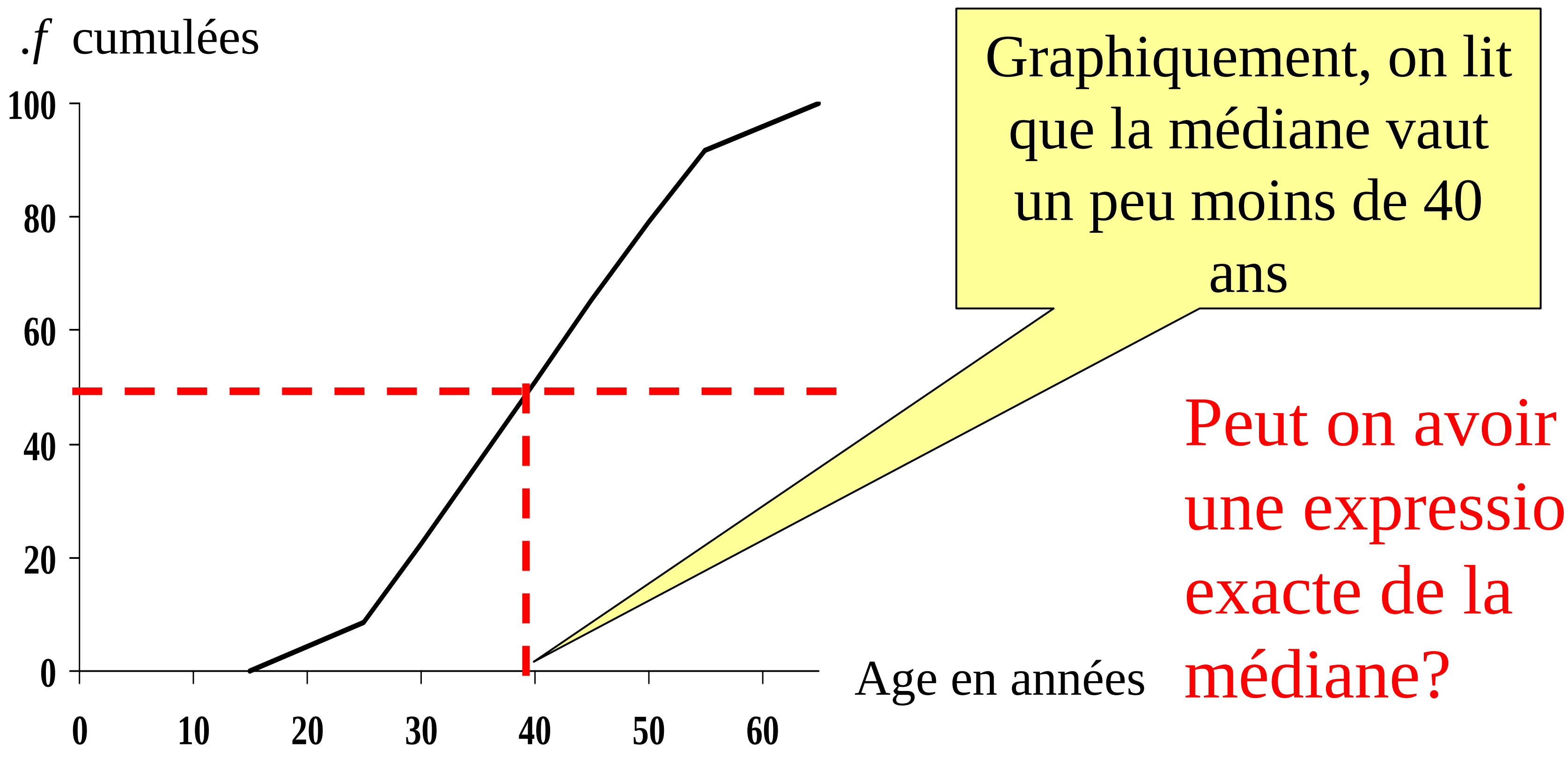
C'est la valeur correspondant à un effectif cumulé de 50% sur le polygone des fréquences cumulées



La médiane

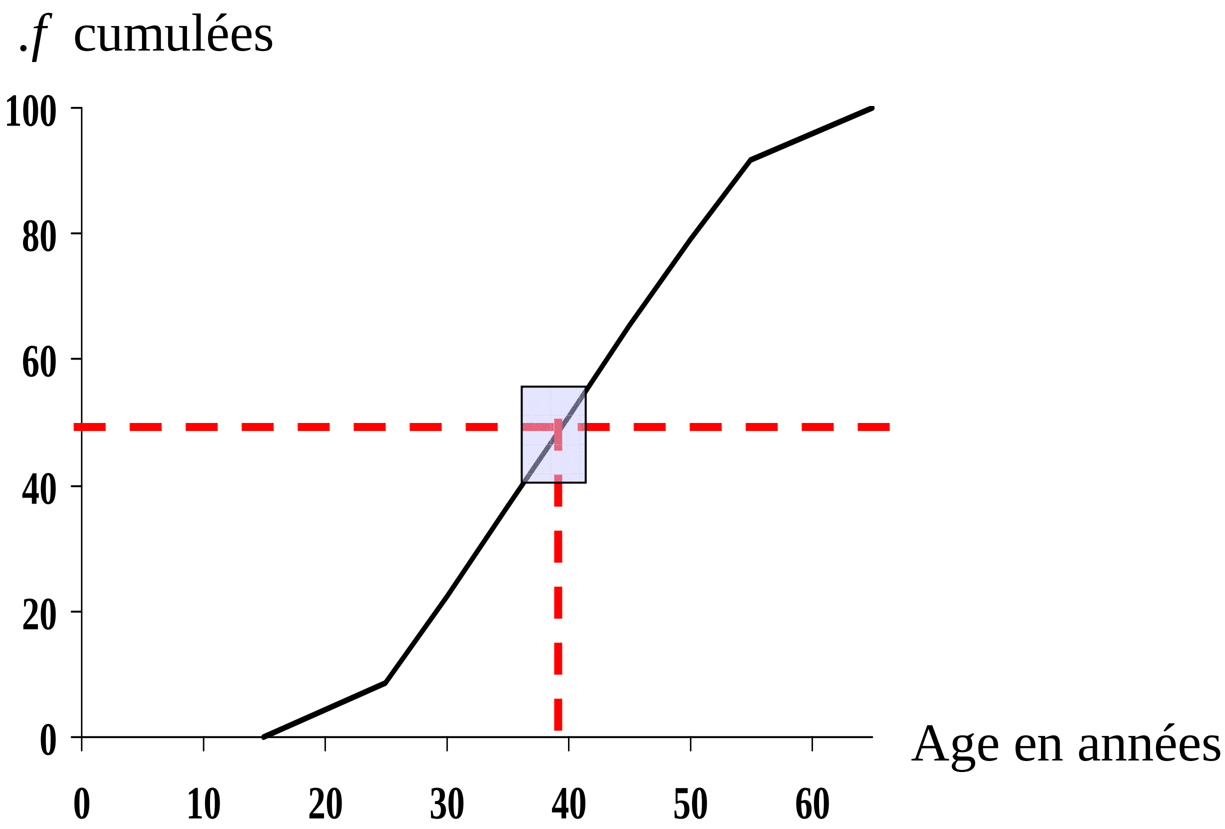
Valeur du caractère qui partage la série statistique en 2 groupes de même fréquence (0.5).

C'est la valeur correspondant à un effectif cumulé de 50% sur le polygone des fréquences cumulées

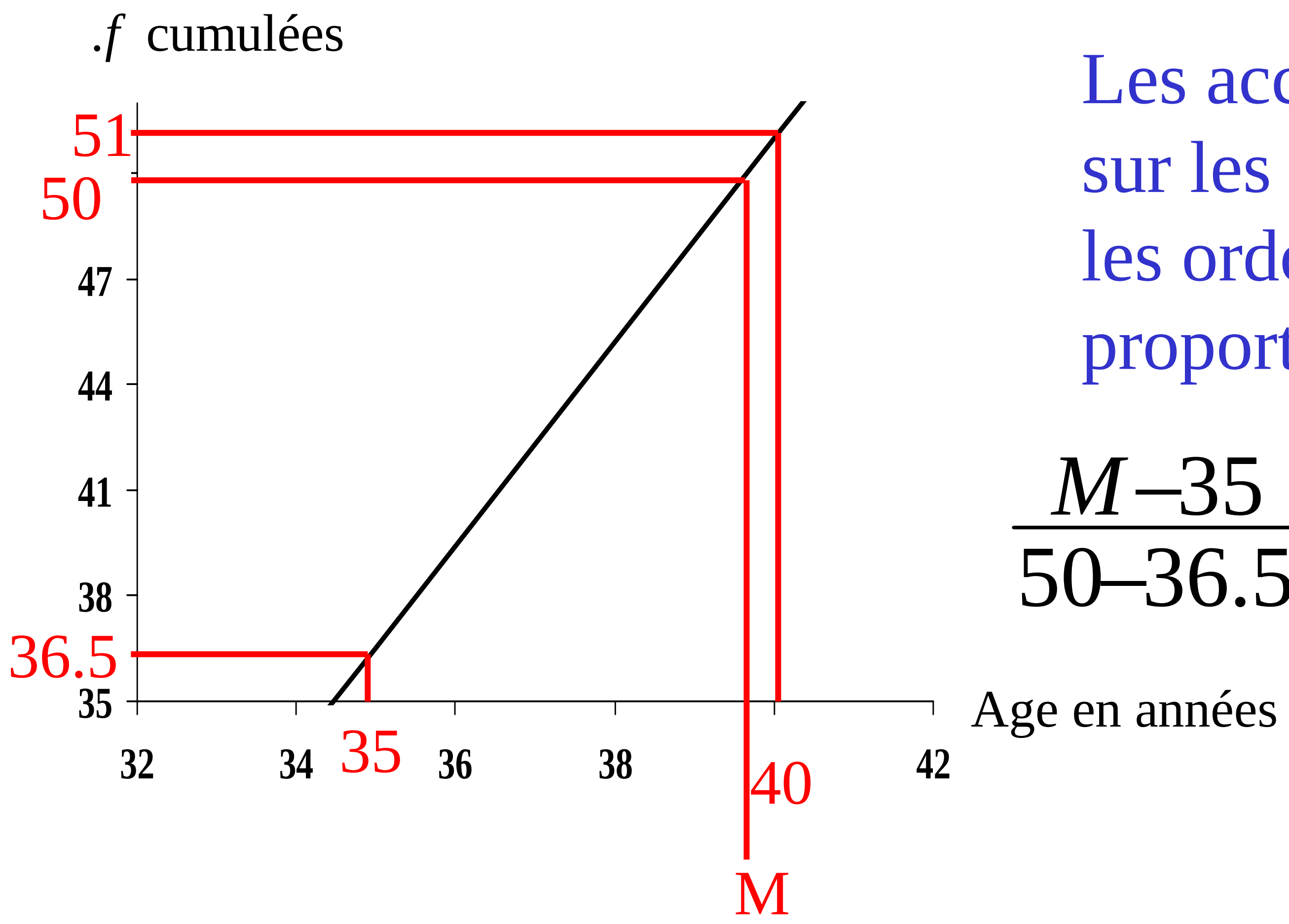


Peut on avoir une expression exacte de la médiane?

Pour avoir la valeur de la médiane on réalise une « interpolation linéaire ».



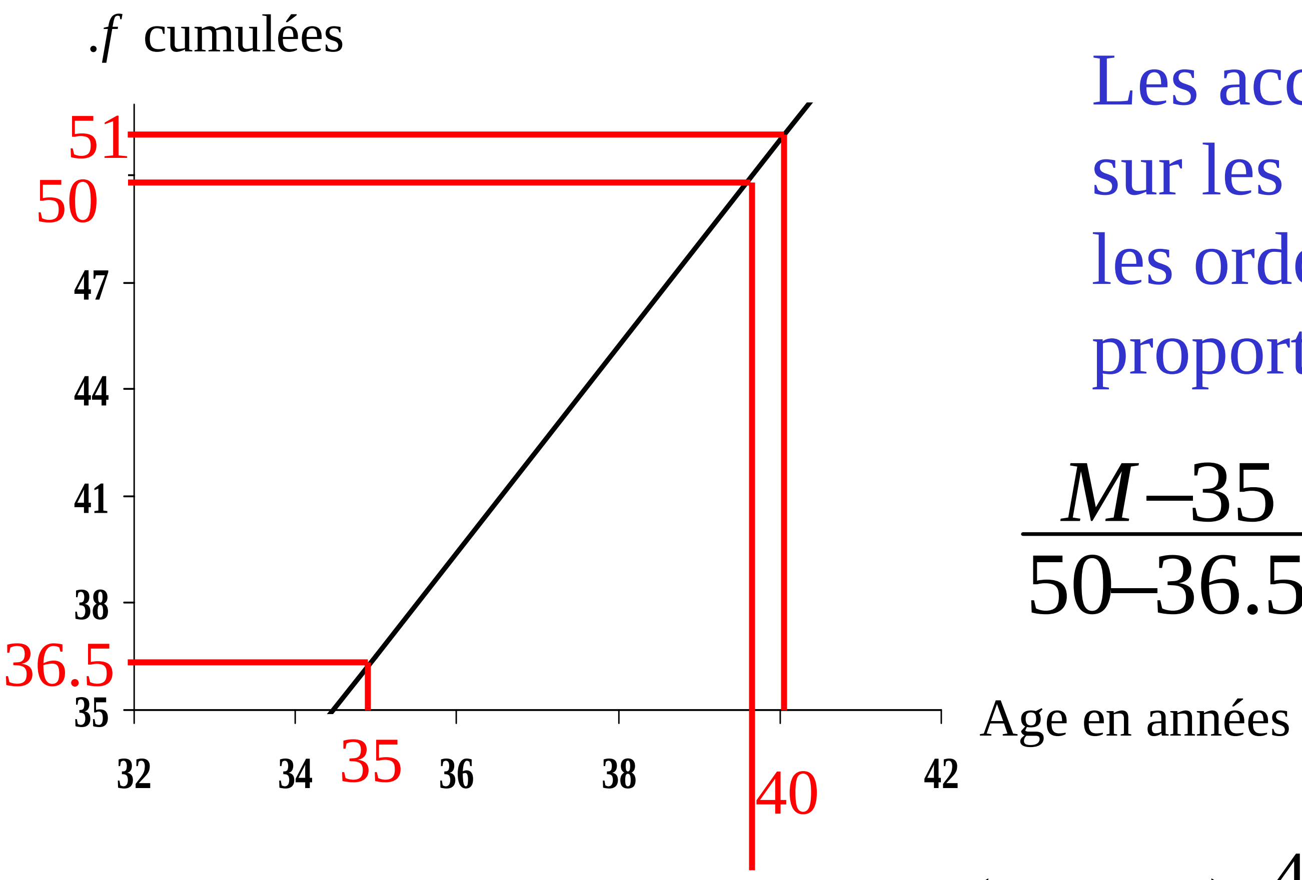
Pour avoir la valeur de la médiane on réalise une « interpolation linéaire ».



Les accroissements sur les abscisses et les ordonnées sont proportionnels

$$\frac{M-35}{50-36.5}=\frac{40-35}{51-36.5}$$

Pour avoir la valeur de la médiane on réalise une « interpolation linéaire ».

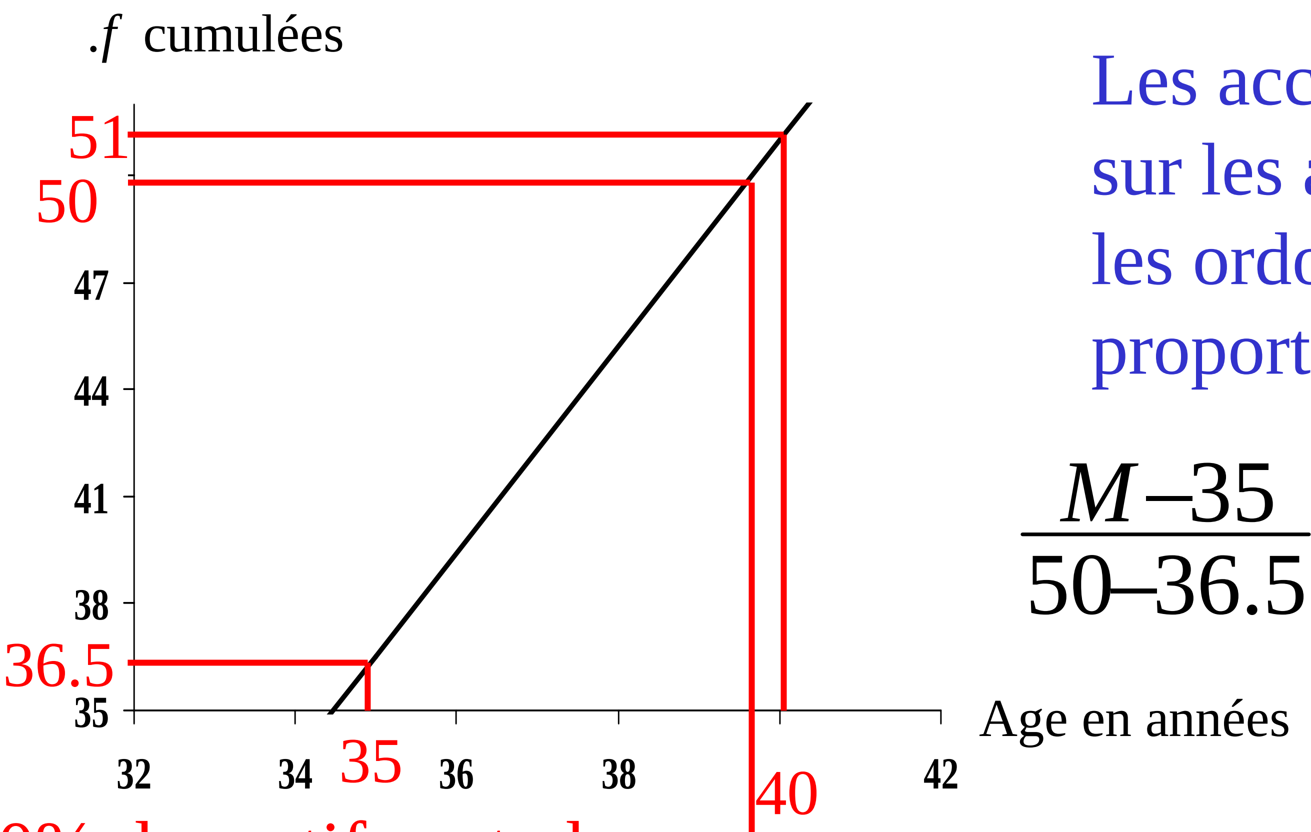


Les accroissements sur les abscisses et les ordonnées sont proportionnels

$$\frac{M-35}{50-36.5}=\frac{40-35}{51-36.5}$$

$$M=35+(50-36.5)\frac{40-35}{51-36.5}\approx 39.7 \text{ (ans)}$$

Pour avoir la valeur de la médiane on réalise une « interpolation linéaire ».



Les accroissements sur les abscisses et les ordonnées sont proportionnels

$$\frac{M-35}{50-36.5}=\frac{40-35}{51-36.5}$$

50% des actifs ont plus de 39.7 ans et 50 % ont moins

$$M=35+(50-36.5)\,\frac{40-35}{51-36.5}\approx 39.7\ \text{(ans)}$$

b. Caractéristiques de dispersion:

Écart absolue, variance, écart-type

Idem caractère discret mais on prend le centre des classes comme valeur représentative

b. Caractéristiques de dispersion:

Écart absolue, variance, écart-type

Idem caractère discret mais on prend le centre des classes comme valeur représentative

$$\overline{x} \approx 40 \text{ (ans)}$$

$.c_i$	Age	Effectif	$.f_i$
20	15-24	2279542	0.086
27	25-29	3628502	0.137
33	30-34	3771554	0.142
37	35-39	3865252	0.146
43	40-44	3770300	0.142
47	45-49	3696642	0.139
53	50-54	3305278	0.125
60	55 et +	2225411	0.084
	Total	26542481	1

b. Caractéristiques de dispersion:

Écart absolue, variance, écart-type

Idem caractère discret mais on prend le centre des classes comme valeur représentative

$$\overline{x} \approx 40 \text{ (ans)}$$

$.c_i$	Age	Effectif	$.f_i$
20	15-24	2279542	0.086
27	25-29	3628502	0.137
33	30-34	3771554	0.142
37	35-39	3865252	0.146
43	40-44	3770300	0.142
47	45-49	3696642	0.139
53	50-54	3305278	0.125
60	55 et +	2225411	0.084
	Total	26542481	1

$$\begin{aligned}
&0.086 * |20-40| \\
&+0.137 * |27-40| \\
&+0.142 * |33-40| \\
&+0.146 * |37-40| \\
&+0.142 * |43-40| \\
&+0.139 * |47-40| \\
&+0.125 * |53-40| \\
&+0.084 * |60-40|
\end{aligned}$$

$$\overline{e_{\overline{x}}} \approx 9.64 \text{ (ans)}$$

$$\begin{aligned}
&0.086 * 20^2 \\
&+0.137 * 27^2 \\
&+0.142 * 33^2 \\
&+0.146 * 37^2 \\
&+0.142 * 43^2 \\
&+0.139 * 47^2 \\
&+0.125 * 53^2 \\
&+0.084 * 60^2
\end{aligned}$$

$$\sigma^2 = 1712 - 40^2 \approx 112 \text{ (ans}^2\text{)}$$

$$\sigma \approx \sqrt{112} \approx 10.6 \text{ (ans)}$$

Le coefficient de variation

$$V = \frac{\sigma}{\bar{x}}$$

Le coefficient de variation

C'est un nombre **SANS UNITE**,
donc plus pratique pour
comparer 2 distributions

$$V = \frac{\sigma}{\bar{x}}$$

Le coefficient de variation

C'est un nombre **SANS UNITE**, donc plus pratique pour comparer 2 distributions

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

Le coefficient de variation

C'est un nombre **SANS UNITE**, donc plus pratique pour comparer 2 distributions

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

Exemple: Prix d'un poisson rouge en Francs à Grenoble

6.5 F   19.5 F   33 F

$$\overline{x_1} \approx 19.7\,(F);\ \ \sigma_1 \approx 10.8\,(F)$$

Prix d'un poisson vert en euros à Grenoble

1 E   3 E   5 E

$$\overline{x_1} \approx 3\,(E);\ \ \sigma_2 \approx 1.63\,(E)$$

Le coefficient de variation

C'est un nombre **SANS UNITE**, donc plus pratique pour comparer 2 distributions

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

Exemple:

Prix d'un poisson rouge en Francs à Grenoble

6.5 F 19.5 F 33 F

$\overline{x}_1 \approx 19.7\,(F);\ \sigma_1 \approx 10.8\,(F)$

$V_1 \approx 0.54$

Prix d'un poisson vert en euros à Grenoble

1 E 3 E 5 E

$\overline{x}_1 \approx 3\,(E);\ \sigma_2 \approx 1.63\,(E)$

$V_2 \approx 0.54$

L'intervalle interquartile

Les quartiles sont les 3 valeurs $Q_1 ; Q_2 ; Q_3$ qui partagent la population en 4 effectifs égaux.

Ce sont les 3 valeurs du caractère correspondant à des effectifs cumulés de 25%, 50% et 75%

L'intervalle interquartile

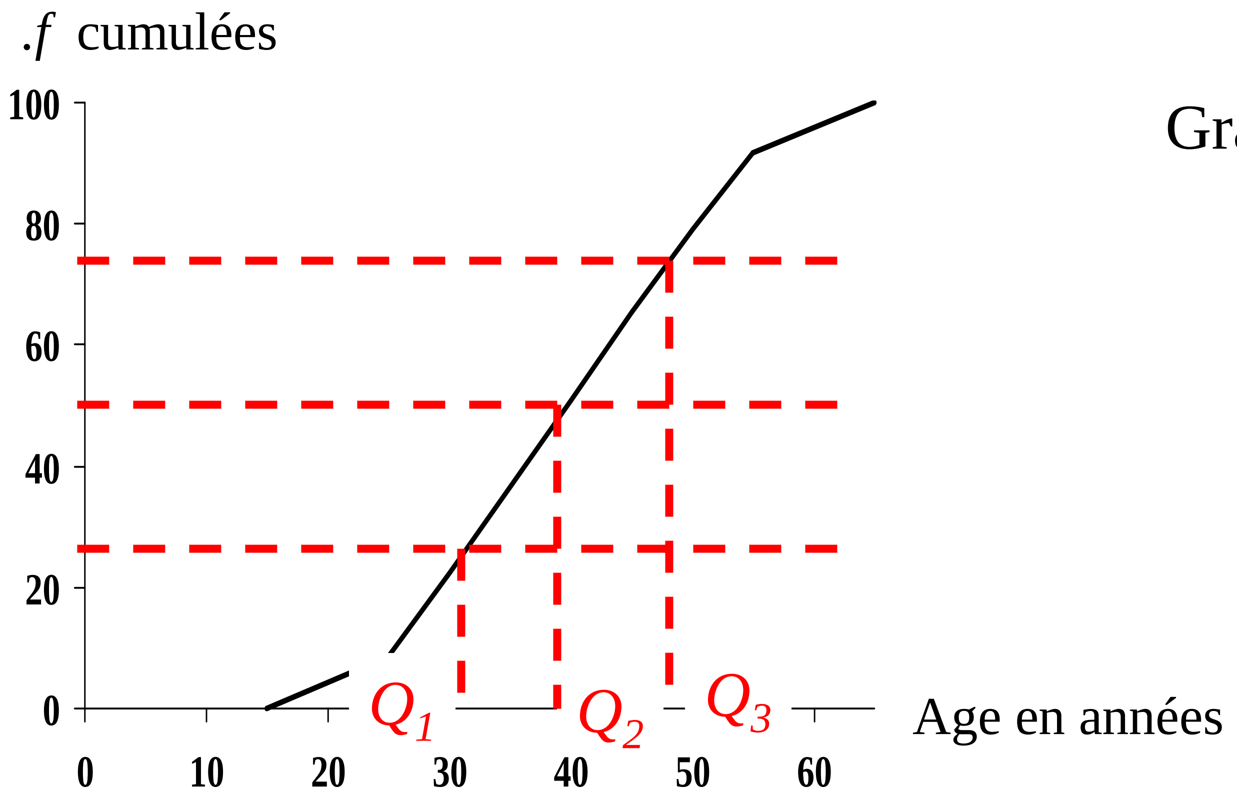
Les quartiles sont les 3 valeurs $Q_1 ; Q_2 ; Q_3$ qui partagent la population en 4 effectifs égaux.

Ce sont les 3 valeurs du caractère correspondant à des effectifs cumulés de 25%, 50% et 75%

L'intervalle interquartile

Les quartiles sont les 3 valeurs $Q_1 ; Q_2 ; Q_3$ qui partagent la population en 4 effectifs égaux.

Ce sont les 3 valeurs du caractère correspondant à des effectifs cumulés de 25%, 50% et 75%



Graphiquement:

$Q_1 \approx 30$ (ans)

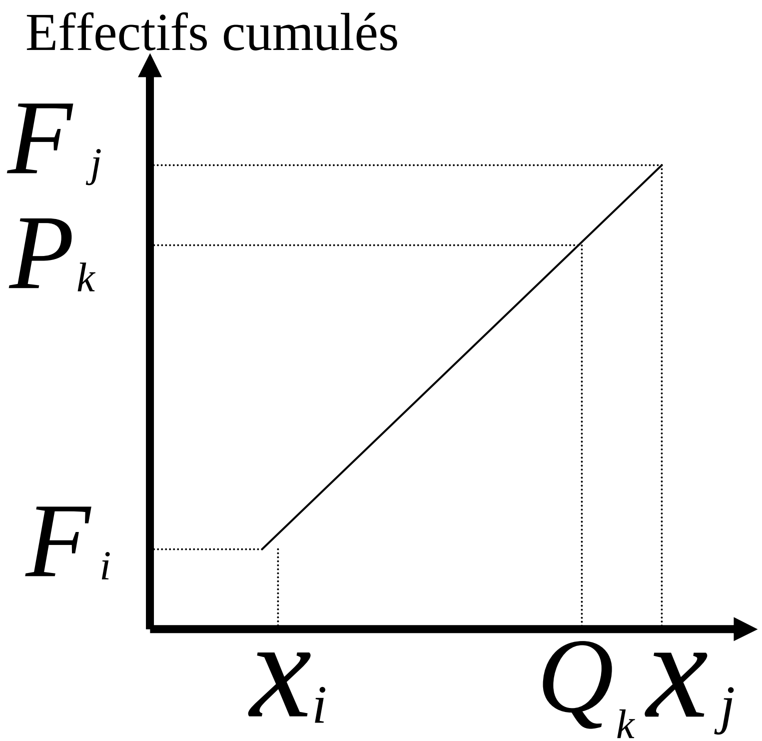
$Q_2 \approx 40$ (ans)

$Q_3 \approx 50$ (ans)

Pour calculer la valeur des quartiles on fait une interpolation linéaire

Pour $k=1,2,3$:

$$Q_k = x_i + (P_k - F_i)\frac{x_j - x_i}{F_j - F_i}$$



$$\begin{cases} P_1 = 25\% \\ P_2 = 50\% \\ P_3 = 75\% \end{cases}$$

Age	Effectif	$.fi$	Cumul
15-24	2279542	0.086	8.6
25-29	3628502	0.137	22.3
30-34	3771554	0.142	36.5
35-39	3865252	0.146	51.0
40-44	3770300	0.142	65.2
45-49	3696642	0.139	79.2
50-54	3305278	0.125	91.6
55 et +	2225411	0.084	100
Total	26542481	1	100

$$Q_1 = 30 + (25 - 22.3)\,\frac{35-30}{36.5-22.3}$$

$$\approx 31\,(\text{ans})$$

$$Q_2 = \text{Me} \approx 39.5\,(\text{ans})$$

$$Q_3 = 45 + (75 - 65.2)\,\frac{50-45}{79.2-65.2}$$

$$\approx 48.5\,(\text{ans})$$

L'intervalle inter-quartile: [Q_1 , Q_3] il contient 50 % de la population et laisse 25% de chaque côté.

L'écart inter-quartile: Q_s est l'amplitude de l'intervalle inter quantile: $Q_s = Q_3 - Q_1$ il mesure la dispersion de la population

L'intervalle inter-quartile: [Q_1 , Q_3] il contient 50 % de la population et laisse 25% de chaque côté.

L'écart inter-quartile: Q_s est l'amplitude de l'intervalle inter quantile: $Q_s = Q_3 - Q_1$ il mesure la dispersion de la population

Exemple:

En France, en 1999, 50 % de la population active a entre 31 et 48.5 ans

$$Q_s = 48.5 - 31 = 17.5 \text{ (ans)}$$

5. Étude d'un couple de caractères

Deux caractères *(X,Y)* pouvant être de nature différente: qualitatif, quantitatif discret ou continu; on note $(x_i)_{i=1..n}$ et $(y_j)_{j=1..m}$ leurs modalités.

Salaire net et âge des livreurs de pizza du restaurant PIPIpizza

Salaires Y Ages X	170-200 Euros	200-230 euros	230-260 euros	
20-22	3	1	0	4
22-24	2	3	0	5
24-26	1	5	1	7
	6	9	1	16

5. Étude d'un couple de caractères

Deux caractères *(X,Y)* pouvant être de nature différente: qualitatif, quantitatif discret ou continu; on note $(x_i)_{i=1..n}$ et $(y_j)_{j=1..m}$ leurs modalités.

Salaire net et âge des livreurs de pizza du restaurant PIPIpizza

3 pers. de 20-22 ans gagnant 170 à 200 euros

Salaires Y Ages X	170-200 Euros	200-230 euros	230-260 euros	
20-22	3	1	0	4
22-24	2	3	0	5
24-26	1	5	1	7
	6	9	1	16

5. Étude d'un couple de caractères

Deux caractères *(X,Y)* pouvant être de nature différente: qualitatif, quantitatif discret ou continu; on note $(x_i)_{i=1..n}$ et $(y_j)_{j=1..m}$ leurs modalités.

Salaire net et âge des livreurs de pizza du restaurant PIPIpizza

3 pers. de 20-22 ans gagnant 170 à 200 euros

9 pers. gagnant entre 200 et 230 euros

Salaires Y Ages X	170-200 Euros	200-230 euros	230-260 euros	
20-22	3	1	0	4
22-24	2	3	0	5
24-26	1	5	1	7
	6	9	1	16

5. Étude d'un couple de caractères

Deux caractères *(X,Y)* pouvant être de nature différente: qualitatif, quantitatif discret ou continu; on note $(x_i)_{i=1..n}$ et $(y_j)_{j=1..m}$ leurs modalités.

Il y a 16 livreurs dans l'entreprise

Salaire net et âge des livreurs de pizza du res[illegible]PIPIpizza

3 pers. de 20-22 ans gagnant 170 à 200 euros

9 pers. gagnant entre 200 et 230 euros

Salaires Y Ages X	170-200 Euros	200-23[illegible] euros	[illegible]-260 [illegible]s	
20-22	3	1	[illegible]	4
22-24	2	3	0	5
24-26	1	5	1	7
	6	9	1	16

5.1 Fréquence relative

F. relative de (x_i, y_j), proportion d'individus présentant la modalité (x_i, y_j) des caractères (X, Y) par rapport à la population totale.

$$f_{i,j} = \frac{n_{i,j}}{N} \qquad \begin{cases} n_{i,j} & \text{Nb individus avec } X = x_i \text{ et } Y = y_i \\ N & \text{Nb totale d'individus} \end{cases}$$

5.1 Fréquence relative

F. relative de (x_i, y_j), proportion d'individus présentant la modalité (x_i, y_j) des caractères (X, Y) par rapport à la population totale.

$$f_{i,j} = \frac{n_{i,j}}{N} \qquad \begin{cases} n_{i,j} & \text{Nb individus avec } X = x_i \text{ et } Y = y_i \\ N & \text{Nb totale d'individus} \end{cases}$$

Propriété: $$\sum_i \sum_j f_{i,j} = 1$$

$$\frac{3}{16} \approx 0.19$$

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4
22-24	2 0.13	3 0.19	0 0	5
24-26	1 0.06	5 0.31	1 0.06	7
	6	9	1	16

$$\frac{3}{16} \approx 0.19$$

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4
22-24	2 0.13	3 0.19	0 0	5
24-26	1 0.06	5 0.31	1 0.06	7
	6	9	1	16

31% des employés ont entre 24 et 26 ans et gagnent entre 200 et 230 euros

5.2 Fréquence marginale

Pour (X,Y) les lois marginales sont:

- La loi de X quelque soit la valeur de Y
- La loi de Y quelque soit la valeur de X

5.2 Fréquence marginale

Pour *(X,Y)* les lois marginales sont:

•La loi de *X* quelque soit la valeur de *Y*

•La loi de *Y* quelque soit la valeur de *X*

Noté: $f_{i,.}$

$f_{.,j}$

5.2 Fréquence marginale

Pour (X,Y) les lois marginales sont:

- La loi de X quelque soit la valeur de Y

- La loi de Y quelque soit la valeur de X

Noté: $f_{i,.}$

$f_{.,j}$

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4 0.25
22-24	2 0.13	3 0.19	0 0	5 0.31
24-26	1 0.06	5 0.31	1 0.06	7 0.44
	6 0.38	9 0.56	1 0.06	16

$$f_{1,.} = \frac{4}{16} \approx 0.25$$

5.2 Fréquence marginale

Pour (X,Y) les lois marginales sont:

- La loi de X quelque soit la valeur de Y
- La loi de Y quelque soit la valeur de X

Noté: $f_{i,.}$

$f_{.,j}$

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4 0.25
22-24	2 0.13	3 0.19	0 0	5 0.31
24-26	1 0.06	5 0.31	1 0.06	7 0.44
	6 0.38	9 0.56	1 0.06	16

$$f_{1,.} = \frac{4}{16} \approx 0.25$$

31% des livreur ont entre 22 et 24 ans

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4 0.25
22-24	2 0.13	3 0.19	0 0	5 0.31
24-26	1 0.06	5 0.31	1 0.06	7 0.44
	6 0.38	9 0.56	1 0.06	16 1

Propriété: $\sum_i f_{i,.} = 1 \qquad \sum_j f_{.,j} = 1$

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4 0.25
22-24	2 0.13	3 0.19	0 0	5 0.31
24-26	1 0.06	5 0.31	1 0.06	7 0.44
	6 0.38	9 0.56	1 0.06	16 1

0.25

+ 0.31

+ 0.44

0.38 + 0.56 + 0.06

<u>Propriété:</u> $$\sum_i f_{i,.} = 1 \qquad \sum_j f_{.,j} = 1$$

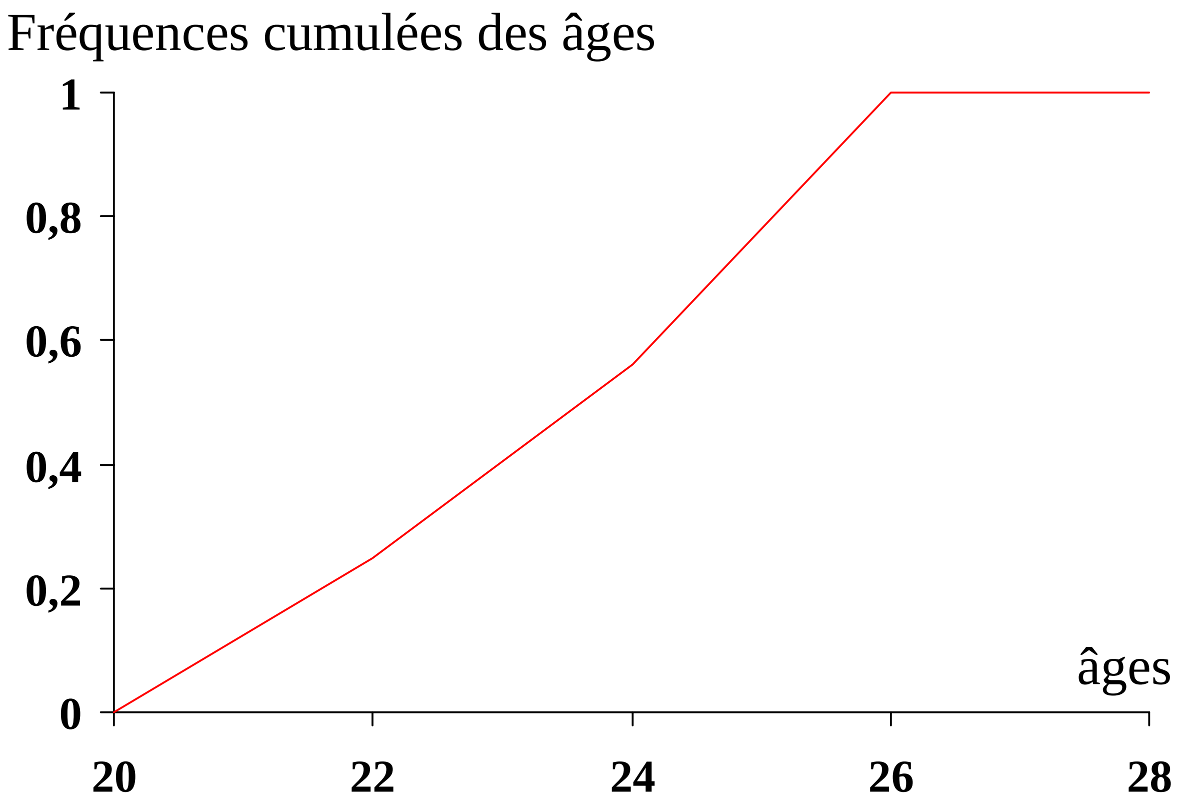
Salaires Y Ages X	170-200 euros		200-230 euros		230-260 euros		
20-22	3 0.19	+	1 0.06	+	0 0	=	4 0.25
22-24	2 0.13	+	3 0.19	+	0 0	=	5 0.31
24-26	1 0.06	+	5 0.31	+	1 0.06	=	7 0.44
	6 0.38		9 0.56		1 0.06		16 1

Propriété: $\sum_j f_{i,j} = f_{\mathrm{i},.}$

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4 0.25
22-24	2 0.13	3 0.19	0 0	5 0.31
24-26	1 0.06	5 0.31	1 0.06	7 0.44
	6 0.38	9 0.56	1 0.06	16 1

<u>Propriété:</u> $\sum_j f_{i,j} = f_{i,.}$ $\sum_i f_{i,j} = f_{.,j}$

Sur les lois marginales, on peut tracer des graphes: de fréquences, fréquences cumulées,…



Sur les lois marginales, on peut calculer des indices centraux et de dispersions.

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.19	1 0.06	0 0	4 0.25
22-24	2 0.13	3 0.19	0 0	5 0.31
24-26	1 0.06	5 0.31	1 0.06	7 0.44
	6 0.38	9 0.56	1 0.06	16 1

Le salaire moyen des livreurs de pizza est de 205.4 euros

185*0.38 +215*0.56+245*0.06 = 205.4 (euros)

5.3 Fréquence conditionnelle

Fréquence conditionnelle de x_i sachant y_i : proportion d'individus présentant la modalité x_i du caractère X par rapport au totale des individus présentant la modalité y_i du caractère Y, notée f_{x_i/y_j}

$$f_{x_i/y_j} = \frac{n_{i,j}}{\sum_i n_{i,j}} \qquad f_{y_j/x_i} = \frac{n_{i,j}}{\sum_j n_{i,j}}$$

Fréquence conditionnelle des âges sachant les salaires

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 $3/6 = 0.5$	1 $1/9 \approx 0.11$	0 $0/1 = 0$	4
22-24	2 $2/6 = 0.3$	3 $3/9 = 0.33$	0 $0/1 = 0$	5
24-26	1 $1/6 \approx 0.17$	5 $5/9 = 0.56$	1 $1/1 = 1$	7
	6	9	1	16

Fréquence conditionnelle des âges sachant les salaires

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 $3/6 = 0.5$	1 $1/9 \approx 0.11$	0 $0/1 = 0$	4
24	2 $2/6 = 0.3$	3 $3/9 = 0.33$	0 $0/1 = 0$	5
24-26	1 $1/6 \approx 0.17$	5 $5/9 = 0.56$	1 $1/1 = 1$	7
	6	9	1	16

Parmi les livreurs gagnant entre 170 et 200 euros, 50% ont entre 20 et 22 ans

Fréquence conditionnelle des âges sachant les salaires

Parmi les livreurs gagnant entre 170 et 200 euros, 50% ont entre 20 et 22 ans

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 $3/6 = 0.5$	1 $1/9 \approx 0.11$	0 $0/1 = 0$	4
	+	+	+	
24	2 $2/6 = 0.3$	3 $3/9 = 0.33$	0 $0/1 = 0$	5
	+	+	+	
24-26	1 $1/6 \approx 0.17$	5 $5/9 = 0.56$	1 $1/1 = 1$	7
	=	=	=	
	6 1	9 1	1 1	16

Fréquence conditionnelle des salaires sachant les âges

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 $3/4 = 0.75$	1 $1/4 \approx 0.25$	0 $0/4 = 0$	4
22-24	2 $2/5 = 0.4$	3 $3/5 = 0.6$	0 $0/5 = 0$	5
24-26	1 $1/7 \approx 0.14$	5 $5/7 = 0.71$	1 $1/7 = 0.14$	7
	6	9	1	16

Fréquence conditionnelle des salaires sachant les âges

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 $3/4 = 0.75$	1 $1/4 \approx 0.25$	0 $0/4 = 0$	4
24	2 $2/5 = 0.4$	3 $3/5 = 0.6$	0 $0/5 = 0$	5
24-26	1 $1/7 \approx 0.14$	5 $5/7 = 0.71$	1 $1/7 = 0.14$	7
	6	9	1	16

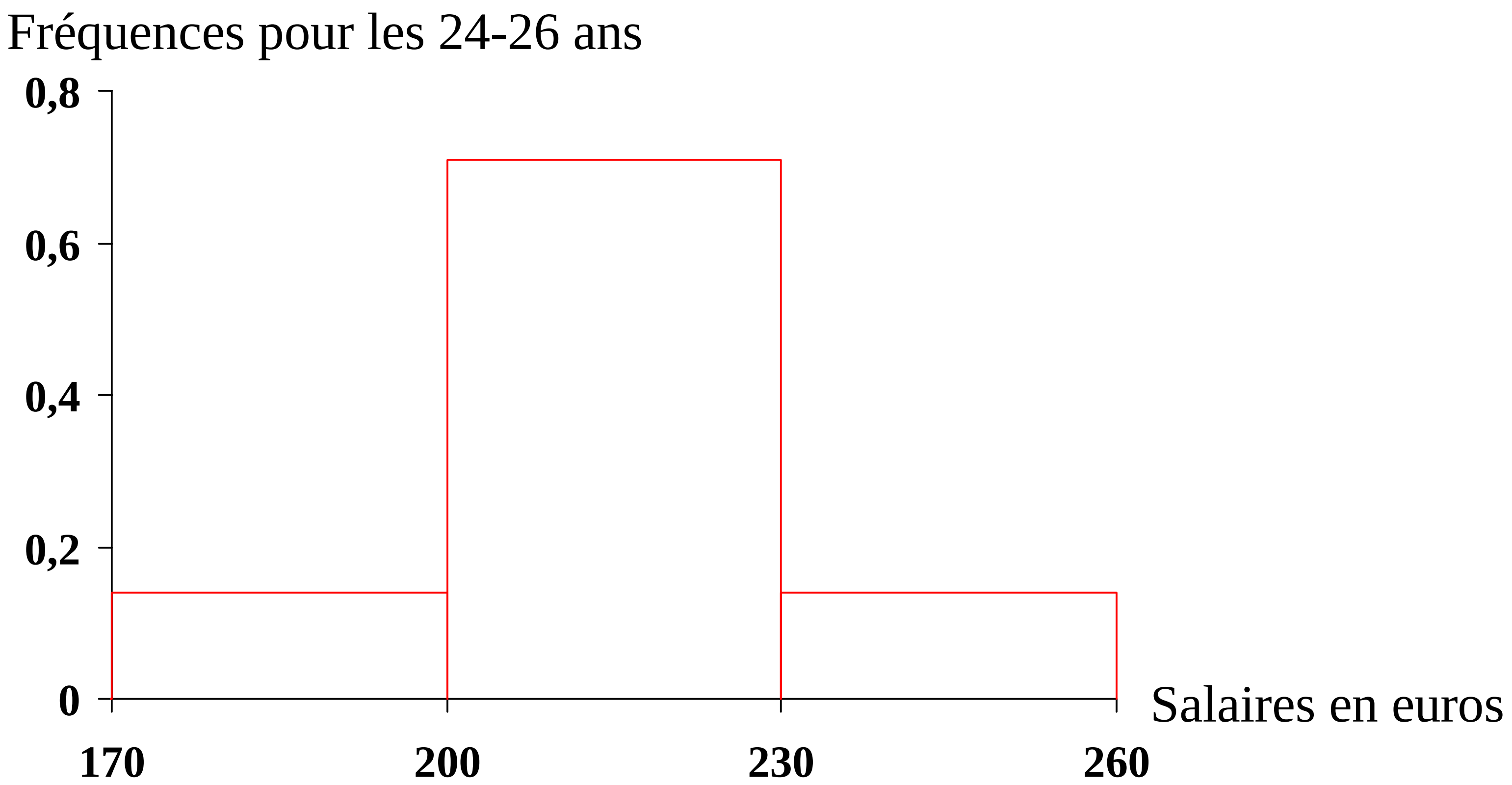
Parmi les livreurs âgés de 20 à 22 ans, 75% gagnent entre 170 et 200 euros

Fréquence conditionnelle des salaires sachant les âges

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 $\frac{3}{4}=0.75$ +	1 $\frac{1}{4}\approx 0.25$ +	0 $\frac{0}{4}=0$ =	4 1
24	2 $\frac{2}{5}=0.4$ +	3 $\frac{3}{5}=0.6$ +	0 $\frac{0}{5}=0$ =	5 1
24-26	1 $\frac{1}{7}\approx 0.14$ +	5 $\frac{5}{7}=0.71$ +	1 $\frac{1}{7}=0.14$ =	7 1
	6	9	1	16

Parmi les livreurs âgés de 20 à 22 ans, 75% gagnent entre 170 et 200 euros

Sur les lois conditionnelles, on peut tracer des graphes: de fréquences, fréquences cumulées



Sur les lois conditionnelles, on peut calculer des indices centraux et de dispersions.

Fréquence conditionnelle des salaires sachant les âges

Salaires Y Ages X	170-200 euros	200-230 euros	230-260 euros	
20-22	3 0.75	1 0.25	0 0	4
22-24	2 0.4	3 0.6	0 0	5
24-26	1 0.14	5 0.71	1 0.14	7
	6	9	1	16

Pour les 22-24 ans:

$$0.4*185+0.6*215+0*245$$

$$=203\ (\text{euros})$$

Parmi les livreurs âgés de 22 à 24 ans, le salaire moyen chez PIPIpizza est de 203 euros

5.3 Indépendance

X est dite indépendante de Y si les variations de Y n'entraînent pas de variation de X

5.3 Indépendance

X est dite indépendante de Y si les variations de Y n'entraînent pas de variation de X

Propriété: Si X est indépendante de Y alors Y est indépendante de X.

5.3 Indépendance

X est dite indépendante de Y si les variations de Y n'entraînent pas de variation de X

Propriété: Si X est indépendante de Y alors Y est indépendante de X.

On dit **X et Y sont indépendants**

5.3 Indépendance

X est dite indépendante de *Y* si les variations de *Y* n'entraînent pas de variation de *X*

Propriété: Si *X* est indépendante de *Y* alors *Y* est indépendante de *X*.

On dit **X et Y sont indépendants**

Les résultats de 2 lancés de dé non pipé sont indépendants!

Propriété:

X et Y sont indépendantes si les fréquences conditionnelles de X sachant Y sont égales aux fréquences marginales de X

Propriété:

X et Y sont indépendantes si les fréquences conditionnelles de X sachant Y sont égales aux fréquences marginales de X

Ou de façon équivalente,

X et Y sont indépendantes si les fréquences conditionnelles de Y sachant X sont égales aux fréquences marginales de Y

Propriété:

X et Y sont indépendantes si les fréquences conditionnelles de X sachant Y sont égales aux fréquences marginales de X

Ou de façon équivalente,

X et Y sont indépendantes si les fréquences conditionnelles de Y sachant X sont égales aux fréquences marginales de Y

Propriété:

Dans le cas ou il y a indépendance entre X et Y, alors dans le tableau de contingence les valeurs des lignes sont proportionnelles et les valeurs des colonnes le sont aussi.

.f sachant âge	170-200 euros	200-230 euros	230-260 euros	.f des classes d'âge
20-22	0.75	0.25	0	0.25
22-24	0.4	0.6	0	0.31
24-26	0.14	0.71	0.14	0.44

Les distribution sont toutes différentes, donc âges et salaires ne sont pas indépendants, il existe une dépendance entre âges et salaires chez PIPIpizza.

5.3 Dépendance totale

X est dit totalement dépendant de Y, si la connaissance de X entraîne la connaissance de Y.

5.3 Dépendance totale

X est dit totalement dépendant de Y, si la connaissance de X entraîne la connaissance de Y.

Dans le tableau de contingence cela ce traduit par le fait qu'il n'y a qu'un effectif non nul par colonne.

5.3 Dépendance totale

X est dit totalement dépendant de Y, si la connaissance de X entraîne la connaissance de Y.

Dans le tableau de contingence cela ce traduit par le fait qu'il n'y a qu'un effectif non nul par colonne.

Si Y est totalement dépendant de X, alors dans le tableau de contingence, il n'y a qu'un effectif non nul par ligne.

5.3 Dépendance totale

X est dit totalement dépendant de Y, si la connaissance de X entraîne la connaissance de Y.

Dans le tableau de contingence cela ce traduit par le fait qu’il n’y a qu’un effectif non nul par colonne.

Ce n’est pas une notion réciproque, contrairement à l’indépendance

Si Y est totalement dépendant de X, alors dans le tableau de contingence, il n’y a qu’un effectif non nul par ligne.

5.3 Dépendance totale

X est dit totalement dépendant de Y, si la connaissance de X entraîne la connaissance de Y.

Dans le tableau de contingence cela ce traduit par le fait qu'il n'y a qu'un effectif non nul par colonne.

Ce n'est pas une notion réciproque, contrairement à l'indépendance

Si Y est totalement dépendant de X, alors dans le tableau de contingence, il n'y a qu'un effectif non nul par ligne.

Il n'y a pas de dépendance totale entre âge et salaire.

<u>Exemple:</u> Y= « Valeur du lancé d'un dé »
X= « gain »

$$X = \begin{cases} 1 \text{ si } Y \text{ est paire} \\ -1 \text{ si } Y \text{ est impaire} \end{cases}$$

X est totalement dépendant de Y

Y n'est pas totalement dépendant de X

Y n'est pas indépendant de X

<u>Exemple:</u> Y= « Valeur du lancé d'un dé »
X= « gain »

$$X = \begin{cases} 1 \text{ si } Y \text{ est paire} \\ -1 \text{ si } Y \text{ est impaire} \end{cases}$$

X est totalement dépendant de Y

Y n'est pas totalement dépendant de X

Y n'est pas indépendant de X

Dans le cas général il n'y a pas indépendance ni dépendance totale: on est entre les deux.

6. Étude d'un couple de caractères sans pondération: régression linéaire

On étudie un couple de caractère X et Y qui soit:

- Quantitatifs
- Sans pondération: chaque modalité du couple (x_i, y_j) apparaît une seule fois

Exemple:

L'entreprise CONCONconserve étudie l'incidence de la pression marketing. Elle enregistre dans 5 zones géographiques, les Ventes y_i (en milliers de boites de conserve) et les Dépenses Publicitaires x_i (en milliers d' euros)

Région *i*	$.y_i$	$.x_i$
1	27	5
2	32	6
3	31	9
4	40	12
5	65	18

6.1 Visualisation de la corrélation

$$X \approx f(Y)\ ?$$

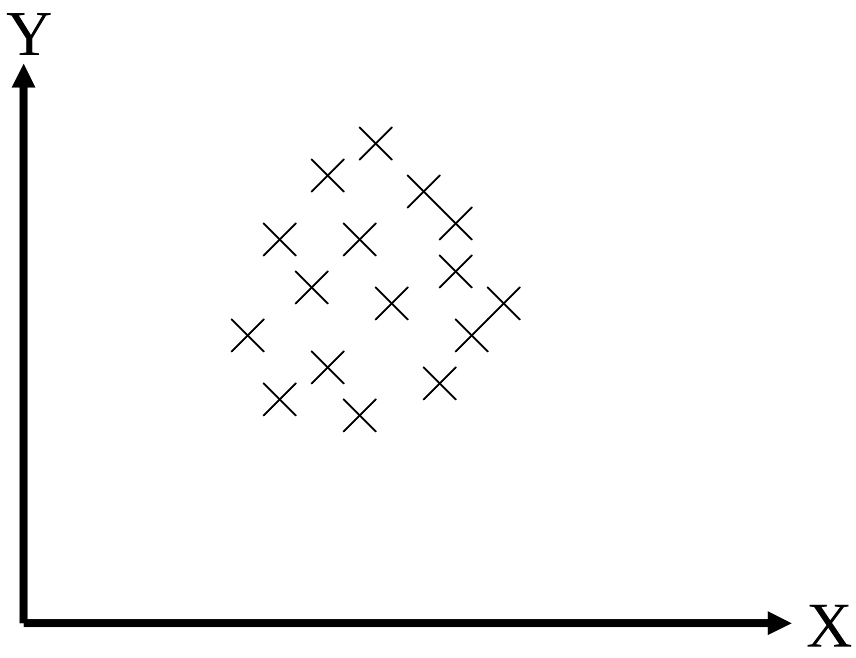
- On représente le nuage de points: X en fonction de Y
- On cherche si il existe une droite ou une courbe qui soit une «bonne approximation » du nuage de points

6.1 Visualisation de la corrélation

$$X \approx f(Y)\ ?$$

- On représente le nuage de points: X en fonction de Y
- On cherche si il existe une droite ou une courbe qui soit une «bonne approximation » du nuage de points

Exemple:

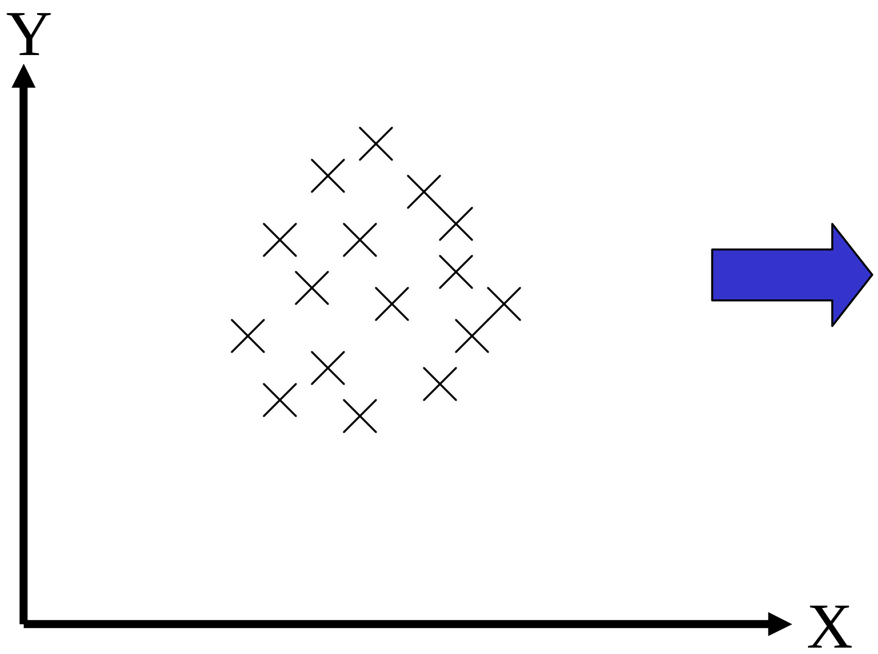


6.1 Visualisation de la corrélation

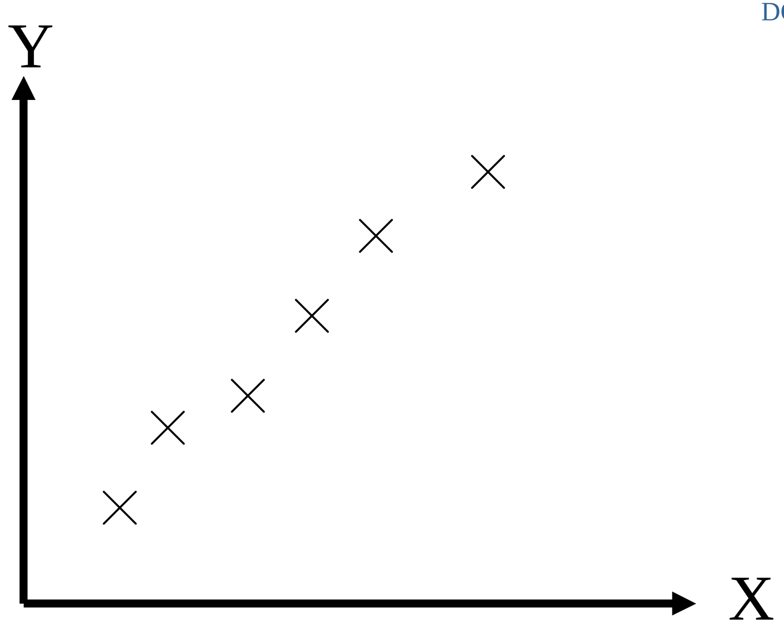
$$X \approx f(Y)\,?$$

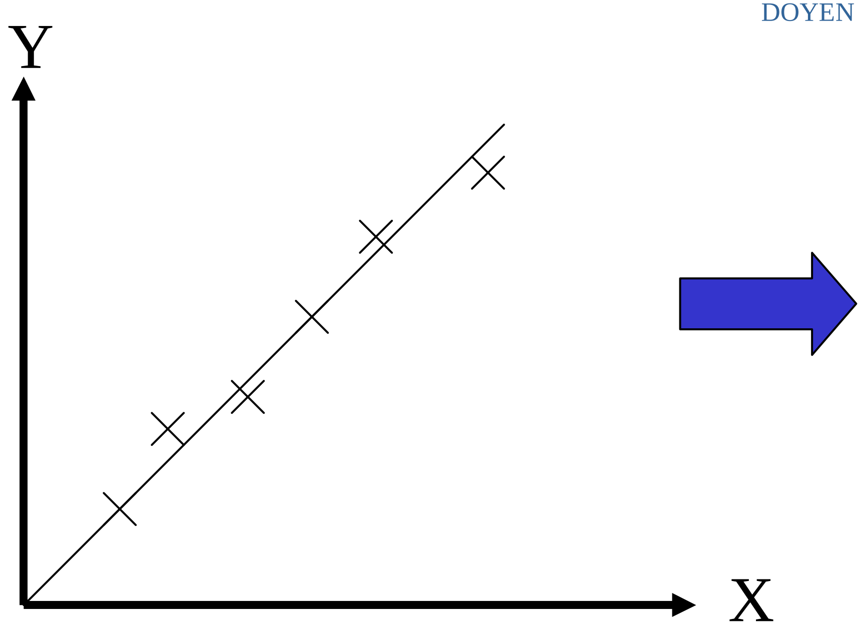
•On représente le nuage de points: X en fonction de Y
•On cherche si il existe une droite ou une courbe qui soit une «bonne approximation » du nuage de points

Exemple:

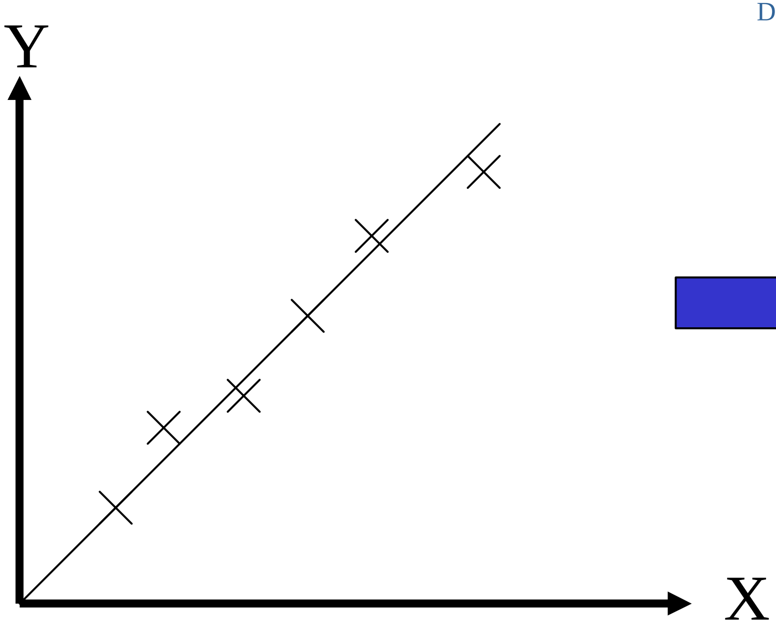


Il n'y a pas de bonne approximation, X et Y semblent indépendants

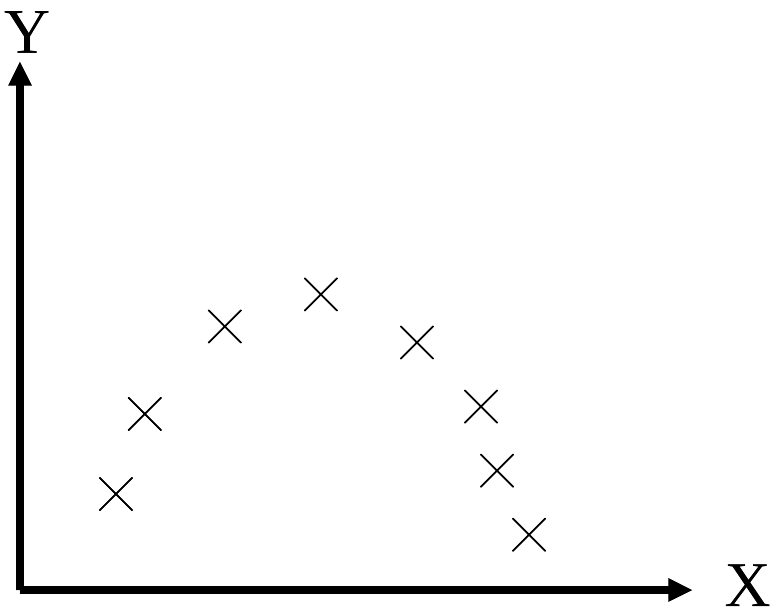

Y
X

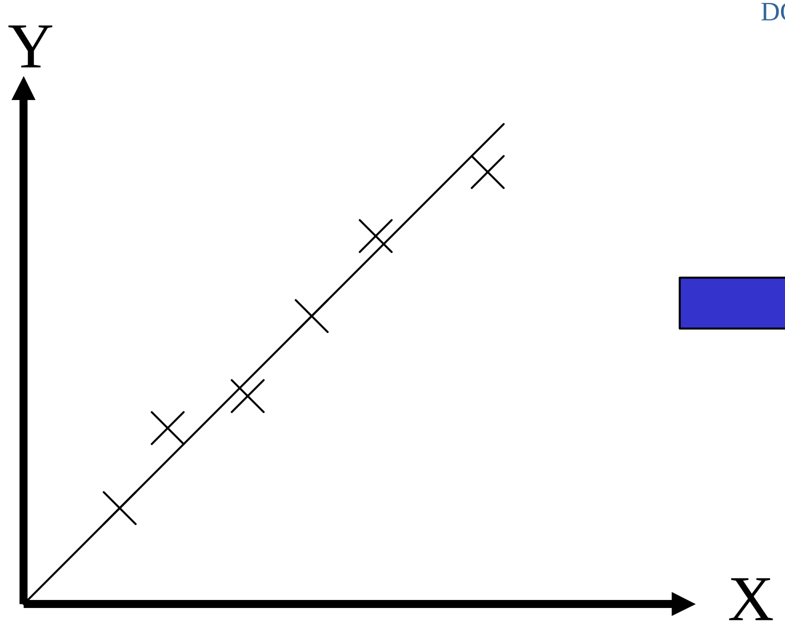


Une droite est une bonne approximation du nuage de points, il existe une relation linéaire entre X et Y.

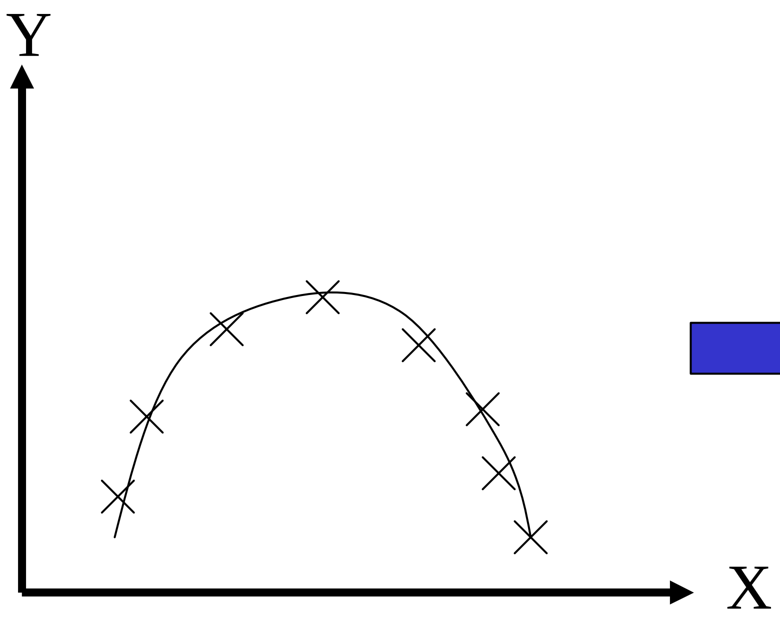


Une droite est une bonne approximation du nuage de points, il existe une relation linéaire entre X et Y.

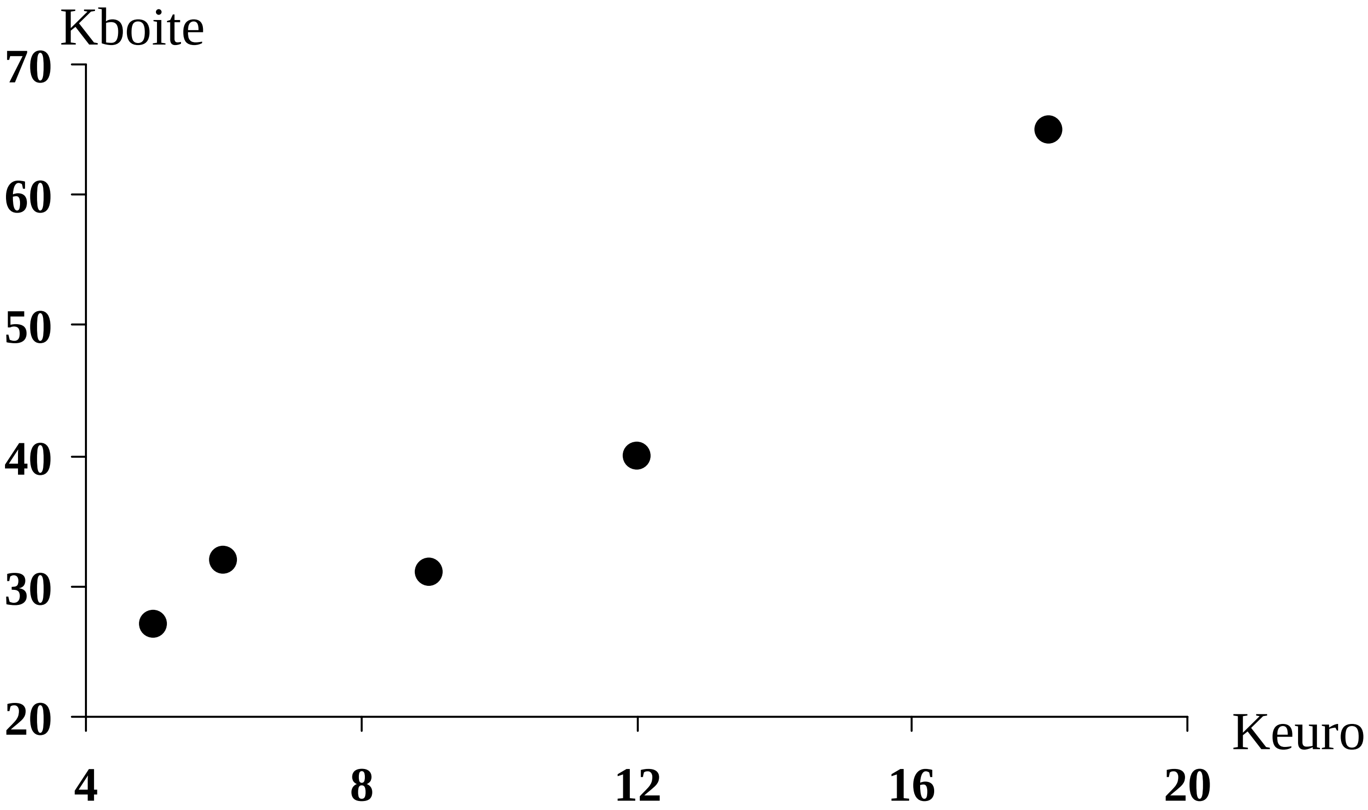


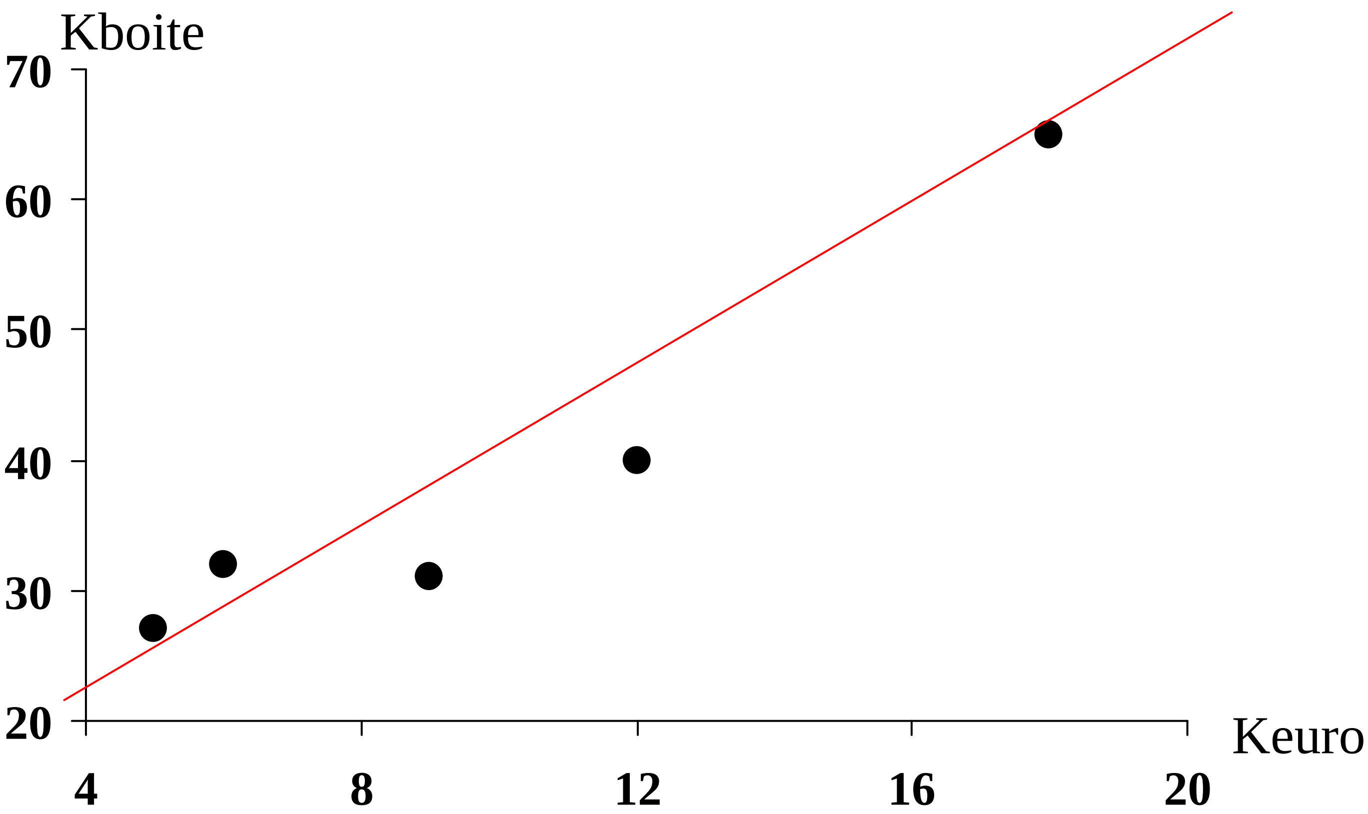


Une droite est une bonne approximation du nuage de points, il existe une relation linéaire entre X et Y.

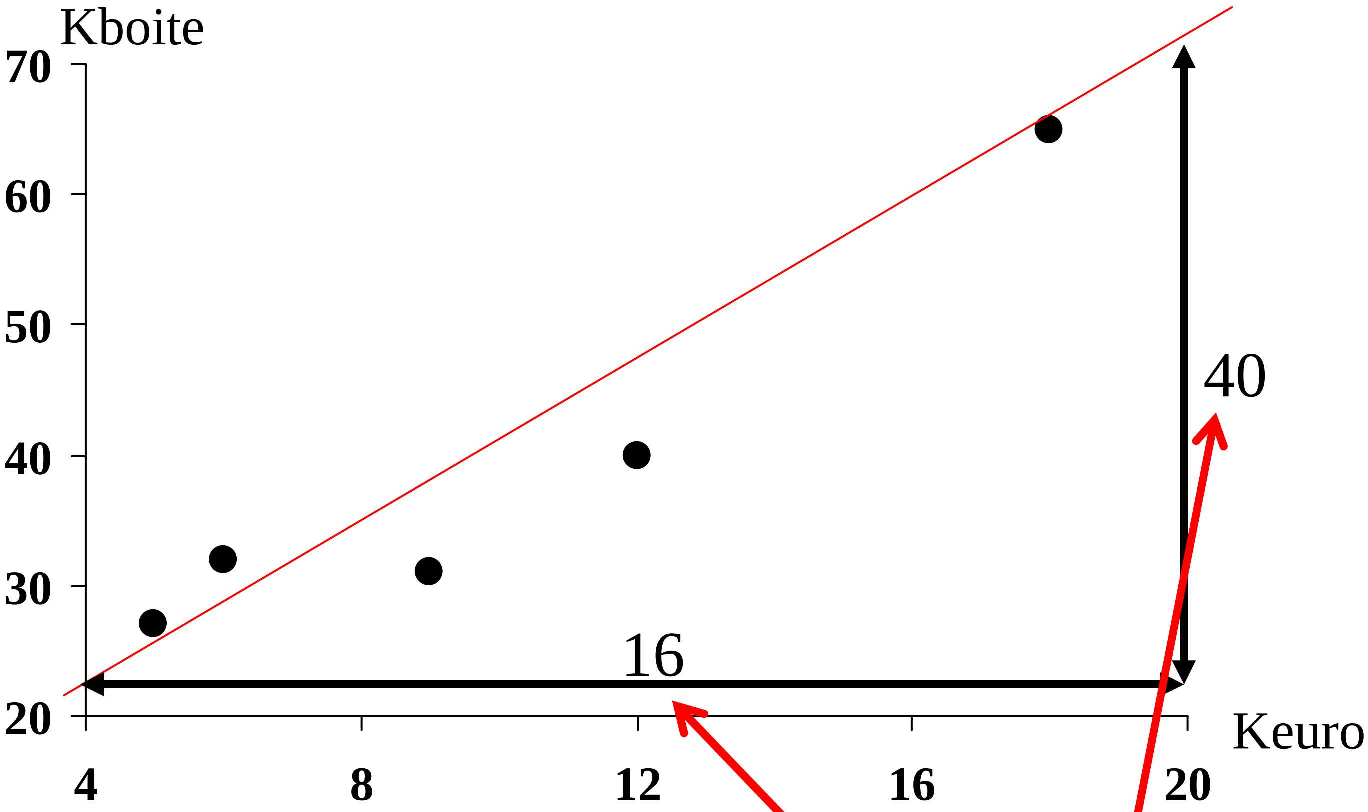


Une courbe est une bonne approximation du nuage de points, il existe une relation curviligne entre X et Y.


Kboite
70
60
50
40
30
20
4
8
12
16
20
Keuro

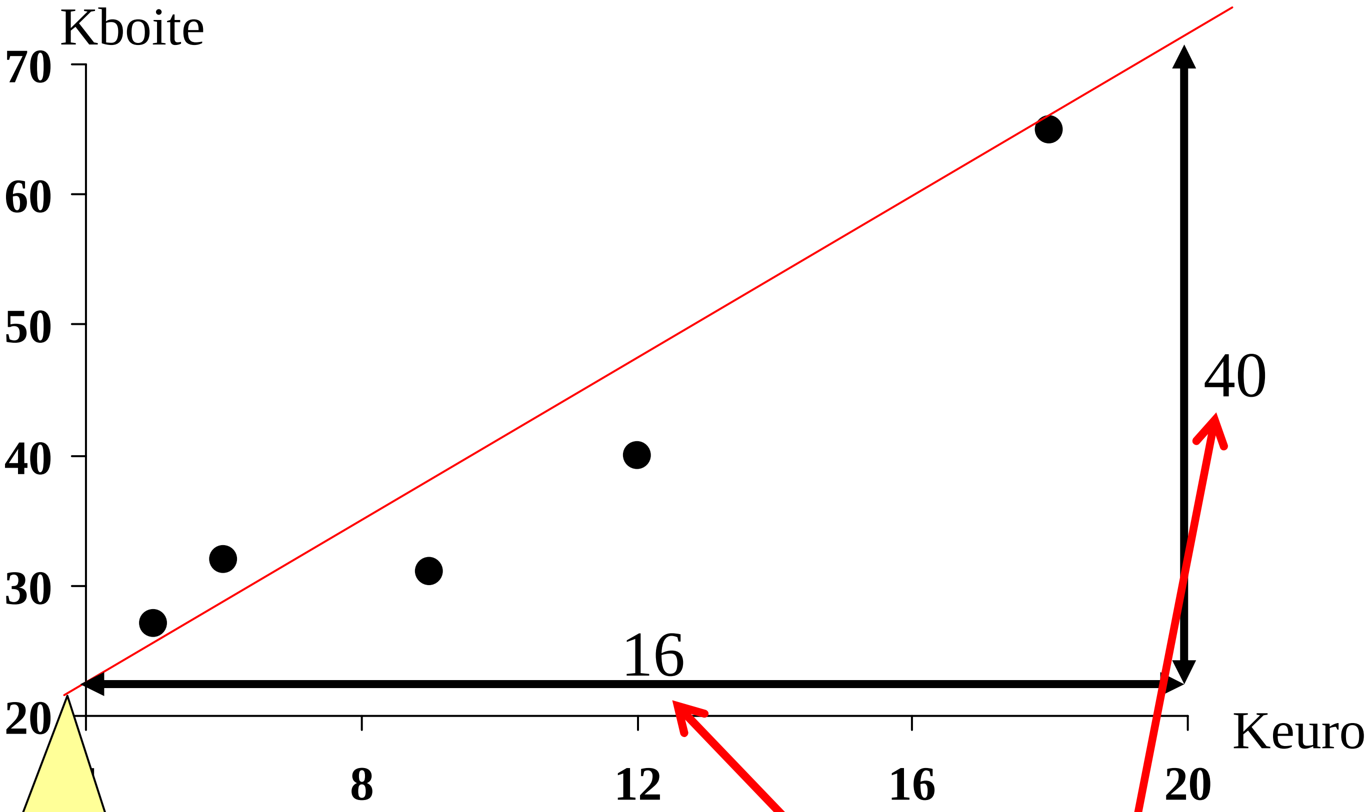


$$Y \approx a * X + b$$



$$Y \approx a^{*}X + b$$

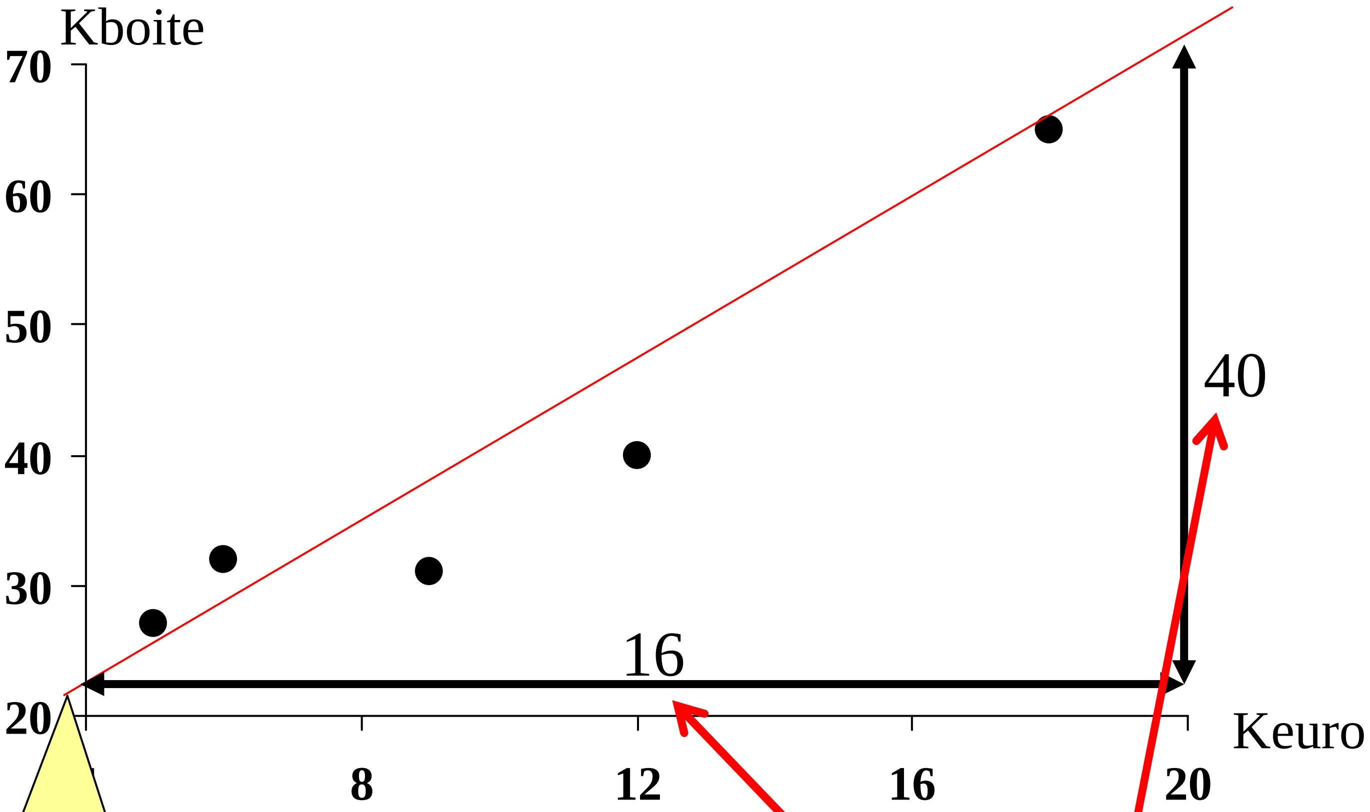
$$a \approx \frac{40}{16} = 2.5\left(\frac{\text{Kboite}}{\text{Keuro}}\right)$$



$$Y \approx a * X + b$$

$$b \approx 20 - 4 * 2.5 = 10 \text{ (Kboite)}$$

$$a \approx \frac{40}{16} = 2.5 \left(\frac{\text{Kboite}}{\text{Keuro}}\right)$$



$$Y \approx a * X + b$$

$$b \approx 20 - 4 * 2.5 = 10 \text{ (Kboite)}$$

$$a \approx \frac{40}{16} = 2.5 \left(\frac{\text{Kboite}}{\text{Keuro}}\right)$$

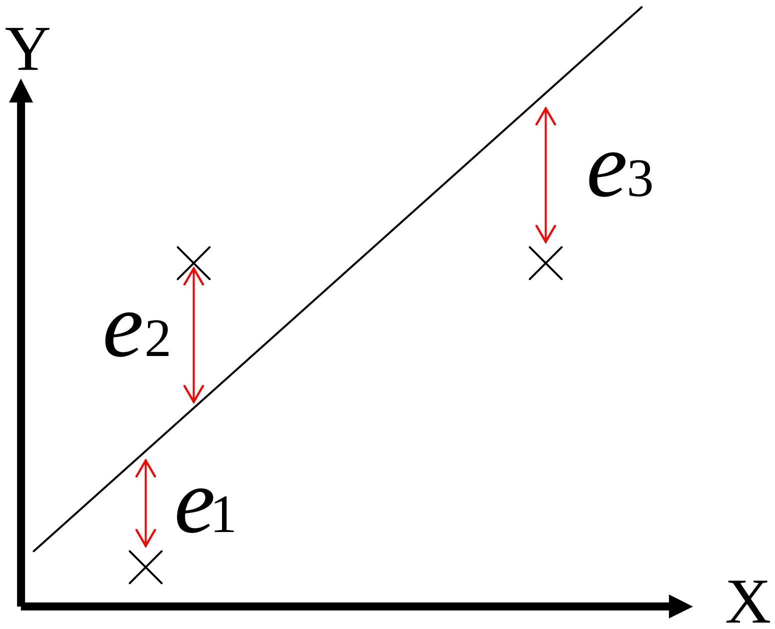
C’est très approximatif!

6.2 L'équation de régression linéaire

Quand l'observation semble être de type linéaire: $Y = a * X + b$

L'objectif est de calculer *a* et *b* de telle sorte que l'on **minimise**:

$$\sum_i e_i^2$$



e_i: Écart entre la droite de régression et la i^ème^ observation

On note: $\overline{x} = \frac{1}{n}\sum_i x_i$ $\overline{y} = \frac{1}{n}\sum_i y_i$

$$V(X) = \frac{1}{n}\sum_i \left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_i x_i^{\,2} - \overline{x}^2$$

$$Cov(X) = \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_i x_i * y_i - \overline{x} * \overline{y}$$

On note: $\overline{x}=\frac{1}{n}\sum_i x_i$ $\overline{y}=\frac{1}{n}\sum_i y_i$

$$V(X)=\frac{1}{n}\sum_i\left(x_i-\overline{x}\right)^2=\frac{1}{n}\sum_i x_i^{\,2}-\overline{x}^{\,2}$$

$$Cov(X)=\frac{1}{n}\sum_i(x_i-\overline{x})(y_i-\overline{y})=\frac{1}{n}\sum_i x_i*y_i-\overline{x}*\overline{y}$$

On a:

$$\boxed{\begin{gathered}a=\frac{Cov(X,Y)}{V(X)}\\ b=\overline{y}-a*\overline{x}\end{gathered}}$$

Région i	$.y_i$	$.x_i$	$.y_i^2$	$.x_i^2$	$.y_i{}^*x_i$
1	27	5	729	25	135
2	32	6	1024	36	192
3	31	9	961	81	279
4	40	12	1600	144	480
5	65	18	4225	324	1170
	195	50	8539	610	2256

Région i	$.y_i$	$.x_i$	$.y_i^2$	$.x_i^2$	$.y_i{}^*x_i$
1	27	5	729	25	135
2	32	6	1024	36	192
3	31	9	961	81	279
4	40	12	1600	144	480
5	65	18	4225	324	1170
	195	50	8539	610	2256

$$\bar{x} = \frac{50}{5} = 10 \text{ (Keuro)}$$

$$\bar{y} = \frac{195}{5} = 39 \text{ (Kboite)}$$

$$V(X) = \frac{610}{5} - 10^2 = 22 \text{ (Keuro)}^2$$

$$Cov(X,Y) = \frac{2256}{5} - 10 * 39 = 61.2 \text{ (Keuro*Kboite)}$$

Région i	$.y_i$	$.x_i$	$.y_i^2$	$.x_i^2$	$.y_i{}^*x_i$
1	27	5	729	25	135
2	32	6	1024	36	192
3	31	9	961	81	279
4	40	12	1600	144	480
5	65	18	4225	324	1170
	195	50	8539	610	2256

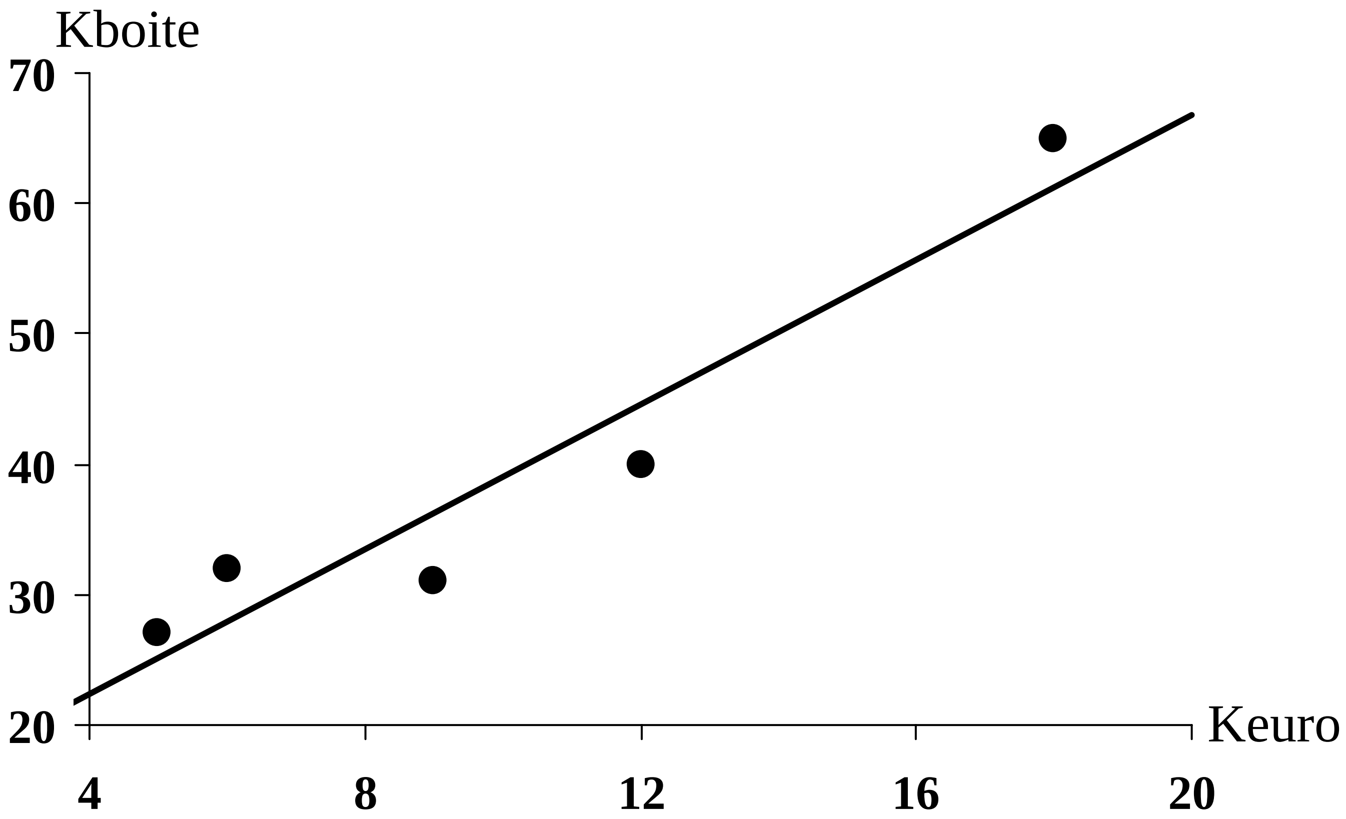
$$\bar{x}=\frac{50}{5}=10 \text{ (Keuro)}$$

$$V(X)=\frac{610}{5}-10^2=22\,(\text{Keuro})^2$$

$$\bar{y}=\frac{195}{5}=39 \text{ (Kboite)}$$

$$Cov(X,Y)=\frac{2256}{5}-10*39=61.2 \text{ (Keuro*Kboite)}$$

$$\boxed{a=\frac{61.2}{22}\approx 2.78\left(\frac{\text{Kboite}}{\text{Keuro}}\right) \qquad b\approx 39-2.78*10=11.2 \text{ (Kboite)}}$$



$$Y \approx 2.78 * X + 11.2$$

6.3 Mesure de la qualité de la régression

Le coefficient de corrélation:

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

6.3 Mesure de la qualité de la régression

Le coefficient de corrélation:

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

Propriétés:

- $-1 \leq r \leq 1$
- $|r|$ proche de 1: corrélation linéaire possible $\left(|r| > 0.86\right)$
- $|r|$ proche de 0: pas de corrélation linéaire

Région *i*	$.y_i$	$.x_i$	$.y_i^2$	$.x_i^2$	$.y_i{}^*x_i$
1	27	5	729	25	135
2	32	6	1024	36	192
3	31	9	961	81	279
4	40	12	1600	144	480
5	65	18	4225	324	1170
	195	50	8539	610	2256

Région i	$.y_i$	$.x_i$	$.y_i^2$	$.x_i^2$	$.y_i{}^*x_i$
1	27	5	729	25	135
2	32	6	1024	36	192
3	31	9	961	81	279
4	40	12	1600	144	480
5	65	18	4225	324	1170
	195	50	8539	610	2256

$$\bar{x}=\frac{50}{5}=10 \text{ (Keuro)} \qquad V(X)=\frac{610}{5}-10^2=22\,(\text{Keuro})^2$$

$$\bar{y}=\frac{195}{5}=39 \text{ (Kboite)} \qquad V(Y)=\frac{8539}{5}-39^2=186.8\,(\text{Kboite})^2$$

$$Cov(X,Y)=\frac{2256}{5}-10*39=61.2 \text{ (Keuro*Kboite)}$$

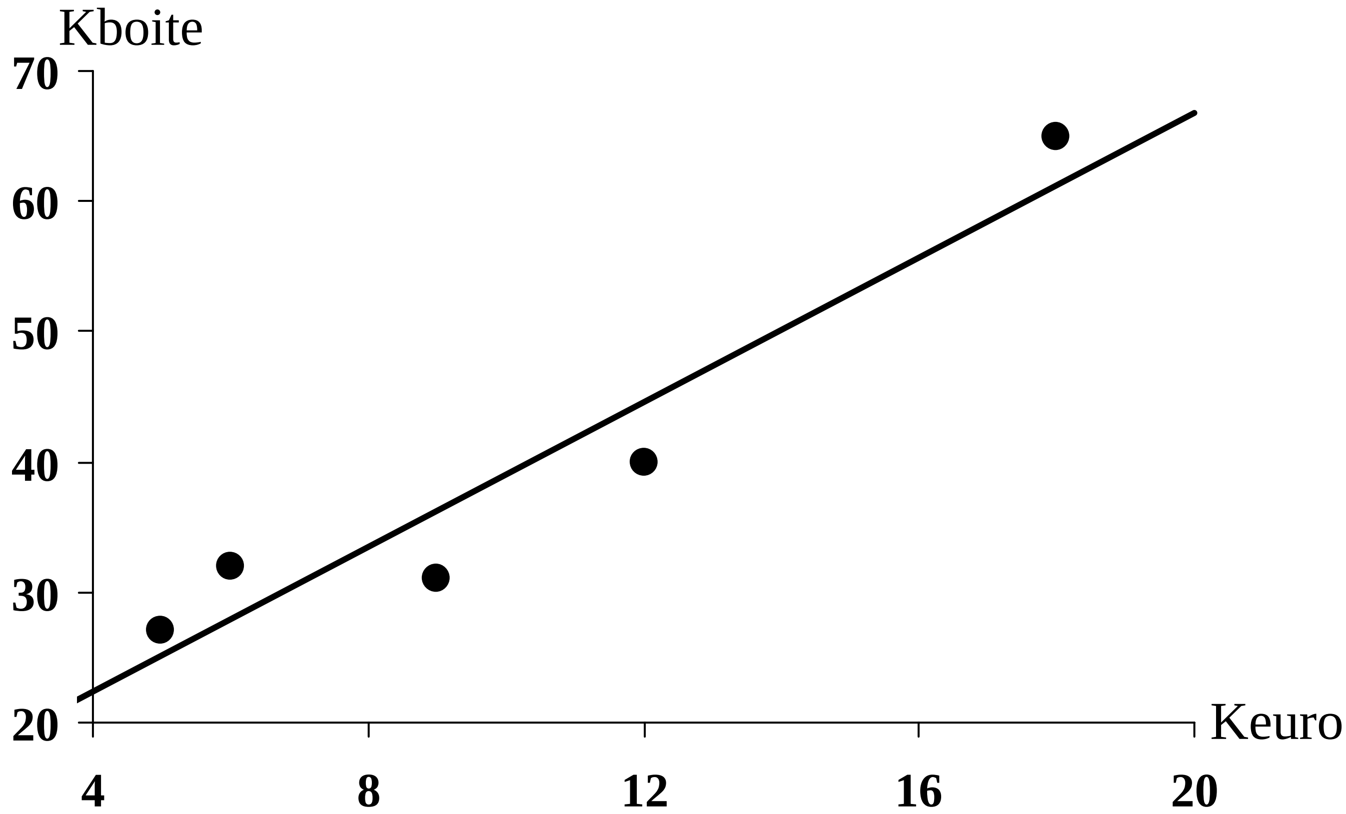
Région i	$.y_i$	$.x_i$	$.y_i^2$	$.x_i^2$	$.y_i{}^*x_i$
1	27	5	729	25	135
2	32	6	1024	36	192
3	31	9	961	81	279
4	40	12	1600	144	480
5	65	18	4225	324	1170
	195	50	8539	610	2256

$$\bar{x}=\frac{50}{5}=10 \text{ (Keuro)} \qquad V(X)=\frac{610}{5}-10^2=22\,(\text{Keuro})^2$$

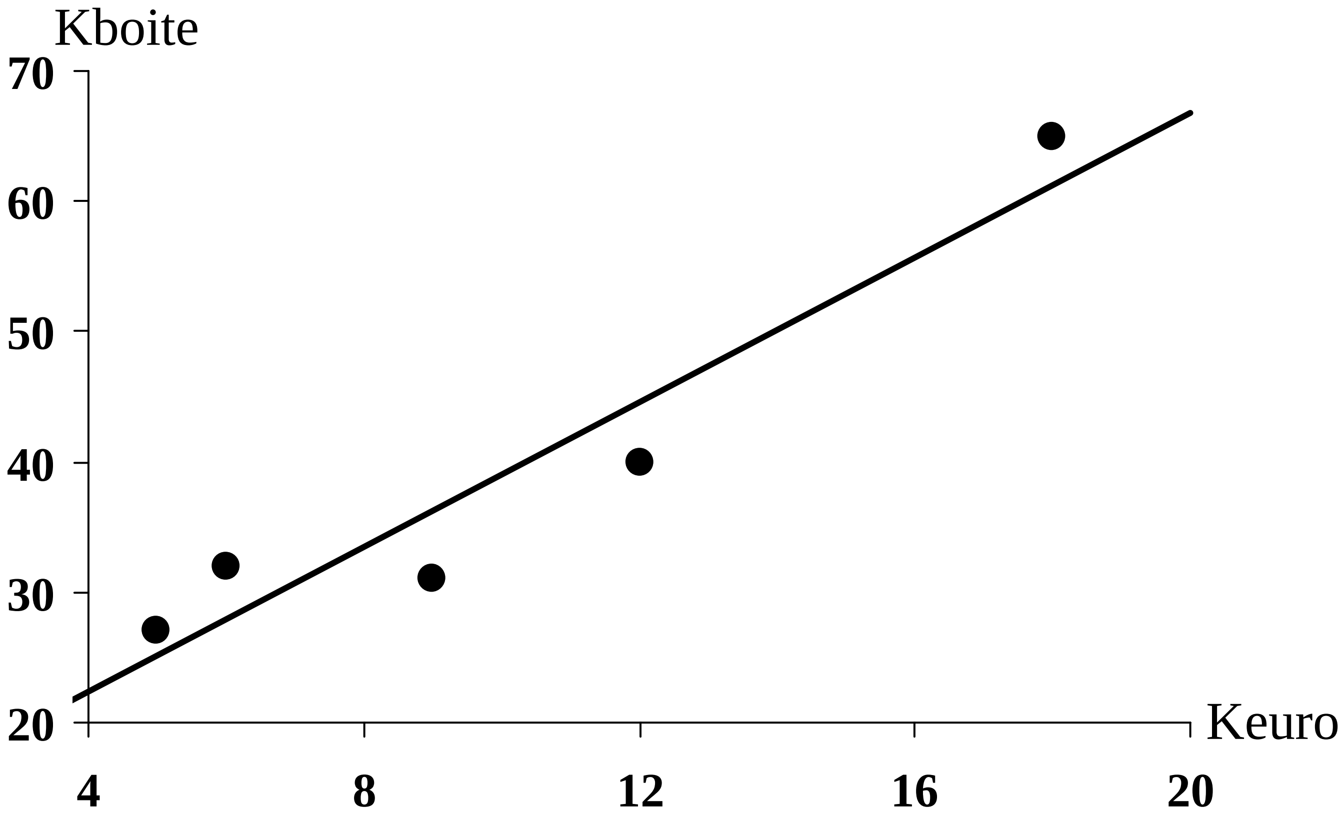
$$\bar{y}=\frac{195}{5}=39 \text{ (Kboite)} \qquad V(Y)=\frac{8539}{5}-39^2=186.8\,(\text{Kboite})^2$$

$$Cov(X,Y)=\frac{2256}{5}-10*39=61.2 \text{ (Keuro*Kboite)}$$

$$\boxed{r\approx\frac{61.2}{\sqrt{22*186.8}}\approx 0.96}$$



$$Y \approx 2.78 * X + 11.2 \qquad r \approx 0.96$$

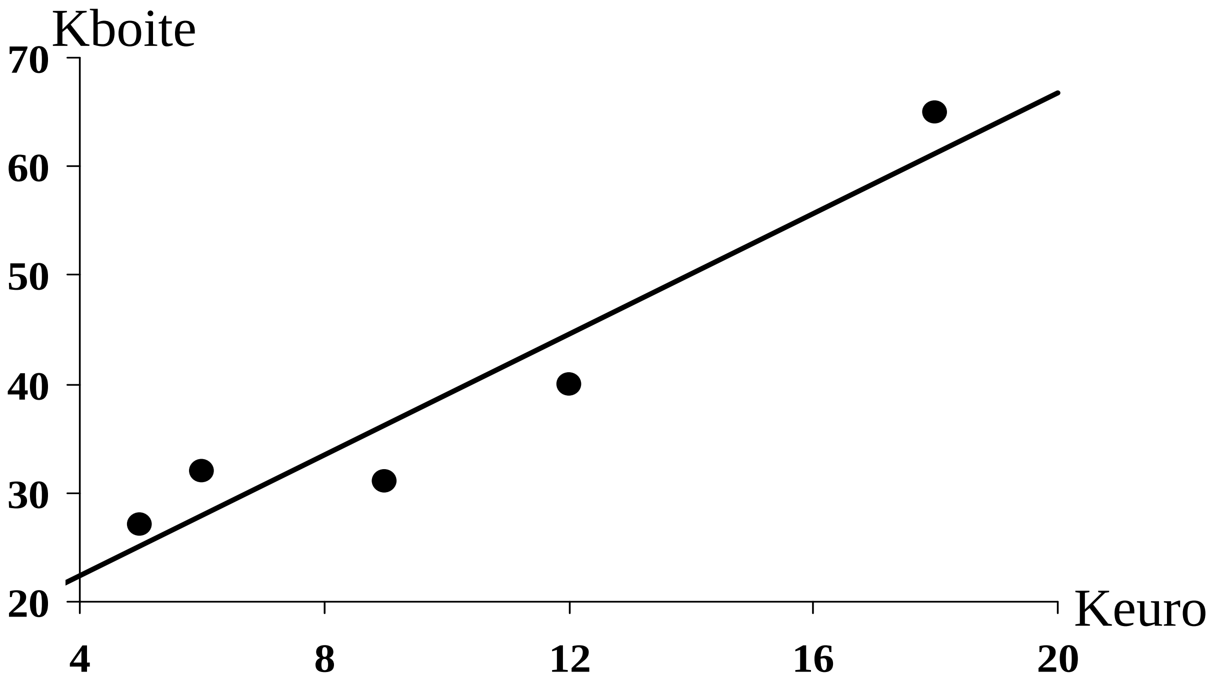


$$Y \approx 2.78 * X + 11.2 \qquad r \approx 0.96$$

La corrélation linéaire des données est forte

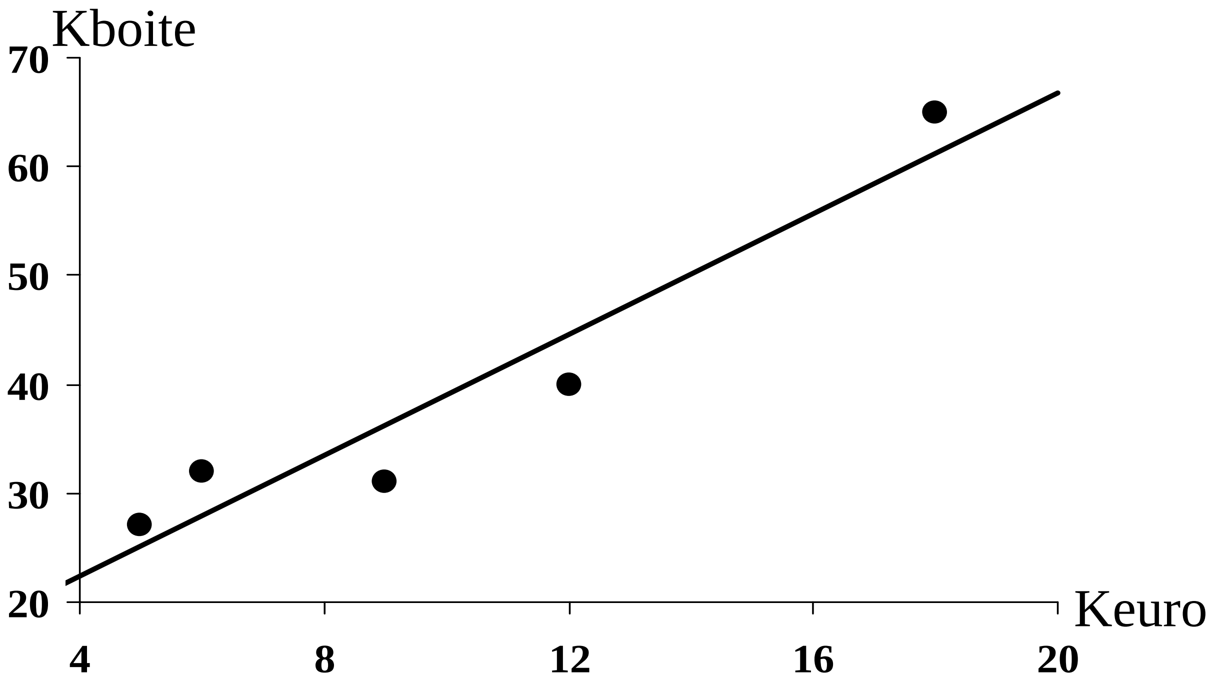
On peut faire de la prévision:

Sur une sixième région on veut vendre $Y=55$ (Kboites), combien faut il dépenser en publicité?



On peut faire de la prévision:

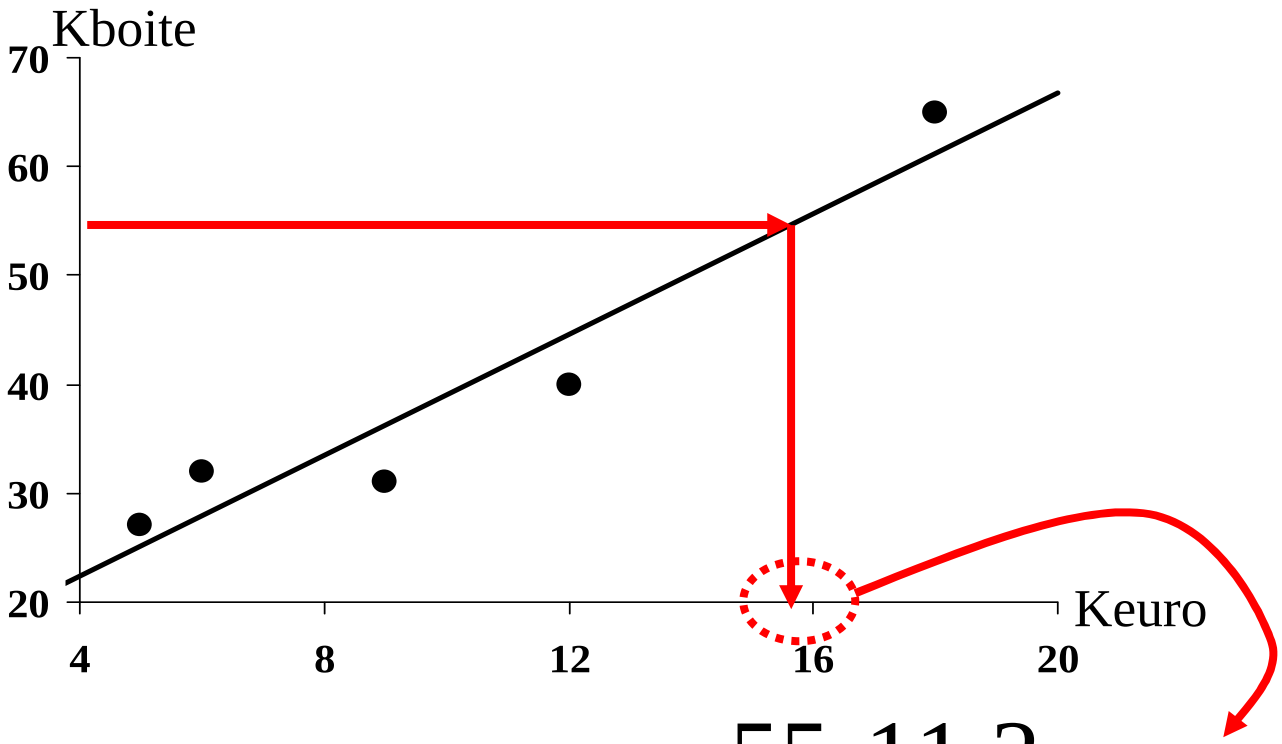
Sur une sixième région on veut vendre *Y=55* (Kboites), combien faut il dépenser en publicité?



$$55 = 2.78 * X + 11.2 \Leftrightarrow X = \frac{55 - 11.2}{2.78} \approx 15.8 \text{ (Keuro)}$$

On peut faire de la prévision:

Sur une sixième région on veut vendre *Y=55* (Kboites), combien faut il dépenser en publicité?



$$55 = 2.78 * X + 11.2 \Leftrightarrow X = \frac{55 - 11.2}{2.78} \approx 15.8 \text{ (Keuro)}$$