

Cours sur statistique descriptive

Données relatives aux employés d'une banque

	A	B	C	D	E	F	G	H	I
1	ID	SEXE	DATENAIS	EDUC	CATEMP	SALDEB	SALACT	TEMPS	EXP
2	1	m	3-févr-1952	15	3	27 000	57 000	98	144
3	2	m	23-mai-1958	16	1	18 750	40 200	98	36
4	3	f	26-juil-1929	12	1	12 000	21 450	98	381
5	4	f	15-avr-1947	8	1	13 200	21 900	98	190
6	5	m	9-févr-1955	15	1	21 000	45 000	98	138
7	6	m	22-août-1958	15	1	13 500	32 100	98	67
8	7	m	26-avr-1956	15	1	18 750	36 000	98	114
9	8	f	6-mai-1966	12	1	9 750	21 900	98	0
10	9	f	23-janv-1946	15	1	13 750	27 900	98	115
11	10	f	13-févr-1946	12	1	13 750	24 000	98	244
12	11	f	7-févr-1950	16	1	16 500	30 300	98	143
13	12	m	11-janv-1966	8	1	10 000	28 350	98	26
14	13	m	17-juil-1960	15	1	13 500	27 750	98	34
15	14	f	26-févr-1949	15	1	16 800	35 100	98	137
16	15	m	29-août-1962	12	1	13 500	27 300	97	66
17	16	m	17-nov-1964	12	1	15 000	40 800	97	24
18	17	m	18-juil-1962	15			46 000	97	48
19	18	m	20-mars-1956	16	3	27 510	103 750	97	70
20	19	m	19-août-1962	12	1	14 250	42 300	97	103
21	20	f	23-janv-1940	12	1	11 550	26 250	97	48
22	21	f	19-févr-1963	16	1	15 000	38 850	97	17
23	22	m	24-sept-1940	12	1	12 750	21 750	97	315
24	23	f	15-mars-1965	15	1	11 100	24 000	97	75
25	24	f	27-mars-1933	12	1	9 000	16 950	97	124
26	25	f	1-juil-1942	15	1	9 000	21 150	97	171
27	26	m	8-nov-1966	15	1	12 600	31 050	96	14
28	27	m	19-mars-1954	19	3	27 480	60 375	96	96

Questions à se poser ?



Comment peut-on rendre plus intelligibles ces données ?



Quelles sont les représentations graphiques requises pour mieux visualiser le comportement de ces variables et quelles interprétations peut-on en faire ?



Existe-t-il des valeurs typiques qui permettraient de résumer l'ensemble des données ?

Objectifs de la partie 1

- Acquérir une culture de base en statistique exploratoire.
- Posséder le sens critique nécessaire à la compréhension de présentations ou de travaux basés sur des études statistiques.
- Savoir choisir les outils adéquats pour le traitement des données, ceci en relation avec **une problématique définie**.
- Pouvoir utiliser de façon adéquate l'outil Python et SPSS.

Qu'est ce que la statistique ?

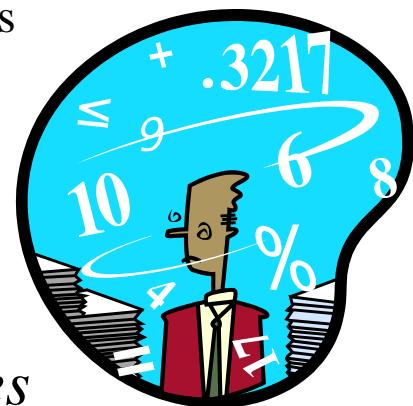
Le mot "statistique" regroupe à la fois un ensemble de données d'observation ainsi que la méthodologie de recueil, de traitement et d'interprétation de ces données.

- **Statistiques (latin « status » état)**

- *Ensemble cohérent de données numériques relatives à un groupe d'individus.*
 - Statistiques démographiques
 - Statistiques annuelles des établissements financiers ou autres
 - Statistiques du chômage

- **Statistique**

- *Ensemble des méthodes qui permettent de rassembler et d'analyser les données numériques*



Rôle et signification de la statistique

- Dans son sens moderne, la **statistique** signifie l'**interprétation** de données numériques, servant de base aux **prises de décisions**.
- La **statistique** comprend également toutes les **méthodes** qui permettent de traduire ces données en **actions**.
- La **différence** principale qui existe entre la **définition** de la **statistique** et l'idée populaire qu'on s'en fait réside dans l'emploi du mot: **singulier** ou **pluriel**.
- En effet, si les **statistiques** sont des **données numériques**, la **statistique**, quant à elle, est une **discipline**.
- L'un des champs d'application de cette discipline est la **statistique descriptive**.
- Celle-ci traite principalement de l'organisation et de la schématisation des données et donne lieu à des illustrations graphiques.
- L '**inférence statistique** est une connaissance essentielle à la prise de décisions en situation d'incertitude.
- Elle permet de tirer des conclusions à partir des données d'échantillon.

Les différentes étapes de toute étude statistique

1. Définition de la problématique de l'étude

2. Collecte des données :

- Simple observation
- Expérimentation : *i.e en provoquant volontairement l'apparition de certains phénomènes contrôlés*

3. Préparation des données

- Consolidation, fiabilité, calcul de nouveaux indicateurs

4. Analyse statistique

- **Analyse "déductive" ou descriptive**
 - a pour but de synthétiser et de présenter les données observées pour que l'on puisse en prendre connaissance facilement : tableaux, graphiques ...
- **Analyse "inductive" ou inférence**
 - permet d'étendre ou de généraliser dans certaines conditions les conclusions obtenues. Cette phase comporte certains risques d'erreur qui peuvent être mesurés en faisant appel à la théorie des probabilités.

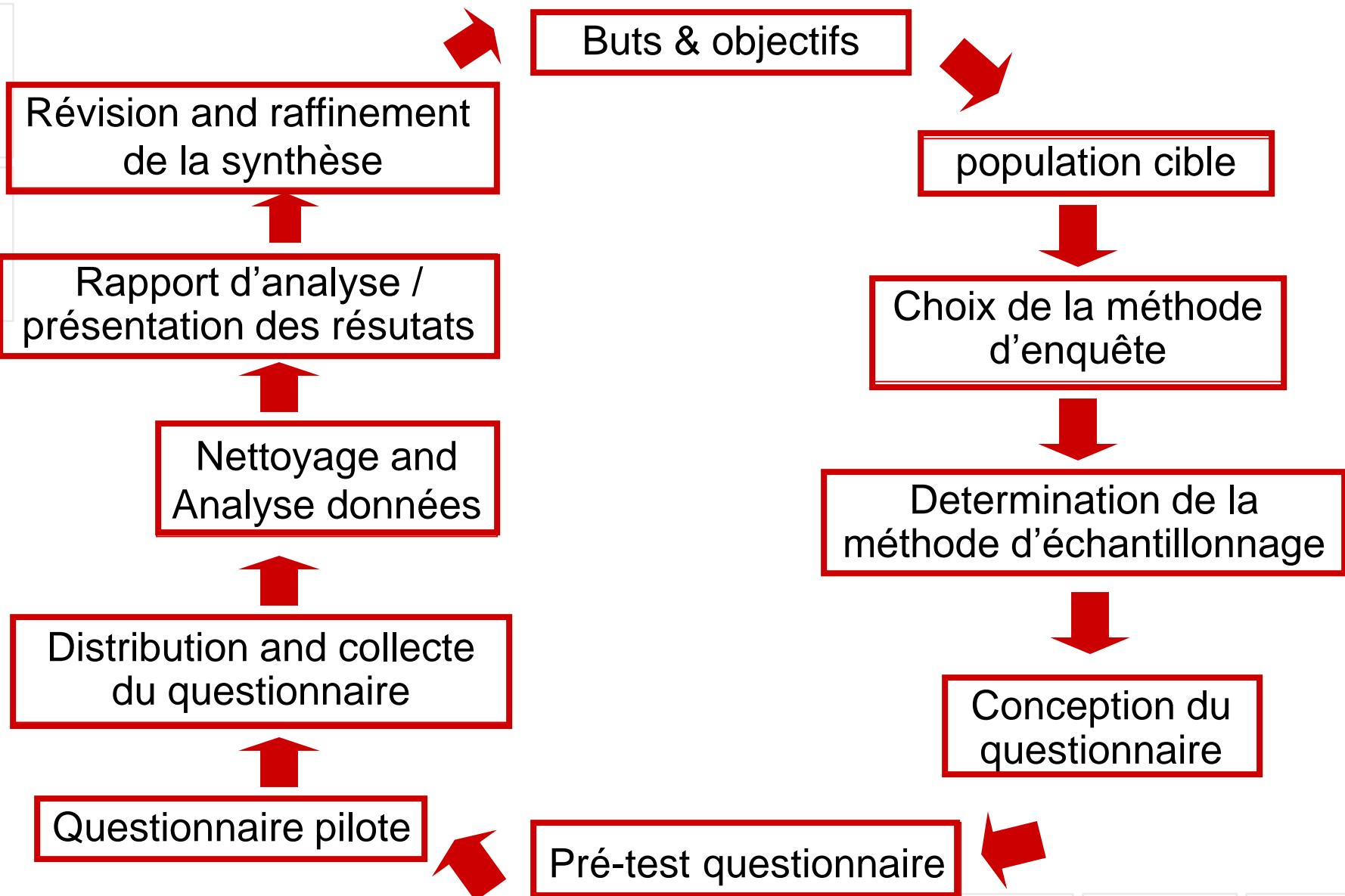
5. Production et diffusion des résultats

Collecte des données

Enquête

- Ensemble des opérations qui ont pour but de collecter de façon organisée des informations relatives à un groupe d'individus ou d'éléments observés dans leur milieu ou leur cadre habituel.
- Lorsque toutes les unités de la population sont observées l'enquête est exhaustive. Elle est encore **appelée recensement**.
- Lorsqu'au contraire, *une partie de la population* est observée, l'enquête est dite partielle ou par échantillonnage. Elle est encore appelée **sondage**. La partie de la population observée constitue l'échantillon.
- L'ensemble des unités auquel on s'intéresse est appelé **population ou univers** ou encore **ensemble statistique**.

Processus d'une enquête



Collecte des données

- Les principaux problèmes qui se posent dans la préparation de l'enquête sont :
 - la définition de l'unité de base et de la population
 - la définition des observations à réaliser
 - le choix d'une méthode de collecte des données
 - le choix d'une méthode d'échantillonnage
 - la détermination de la taille de l'échantillon

Collecte des données

Quelques méthodes d'échantillonnage

- **Échantillonnage aléatoire simple** (*avec ou sans remise*)
le tirage sans remise est le plus fréquent et donne des estimations plus précises pour une même taille d'échantillon.
- **Échantillonnage stratifié ou par grappes**
 - *A utiliser quand la population est très hétérogène et que l'on souhaite s'assurer que ses différentes composantes seront toutes bien représentées. La stratification peut apporter un gain de précision important par rapport à un échantillonnage aléatoire simple.*
- **Échantillonnage à deux ou plusieurs niveaux**
 - Tirage au sort des familles
 - Puis tirage au sort dans chaque famille de la personne enquêtée.
- **Méthode des quotas (quota)** largement utilisée dans les sondages d'opinion.

Toutes les méthodes nécessitent une base d'échantillonnage :
on suppose que l'on dispose d'une liste de toutes les unités qui constituent la population

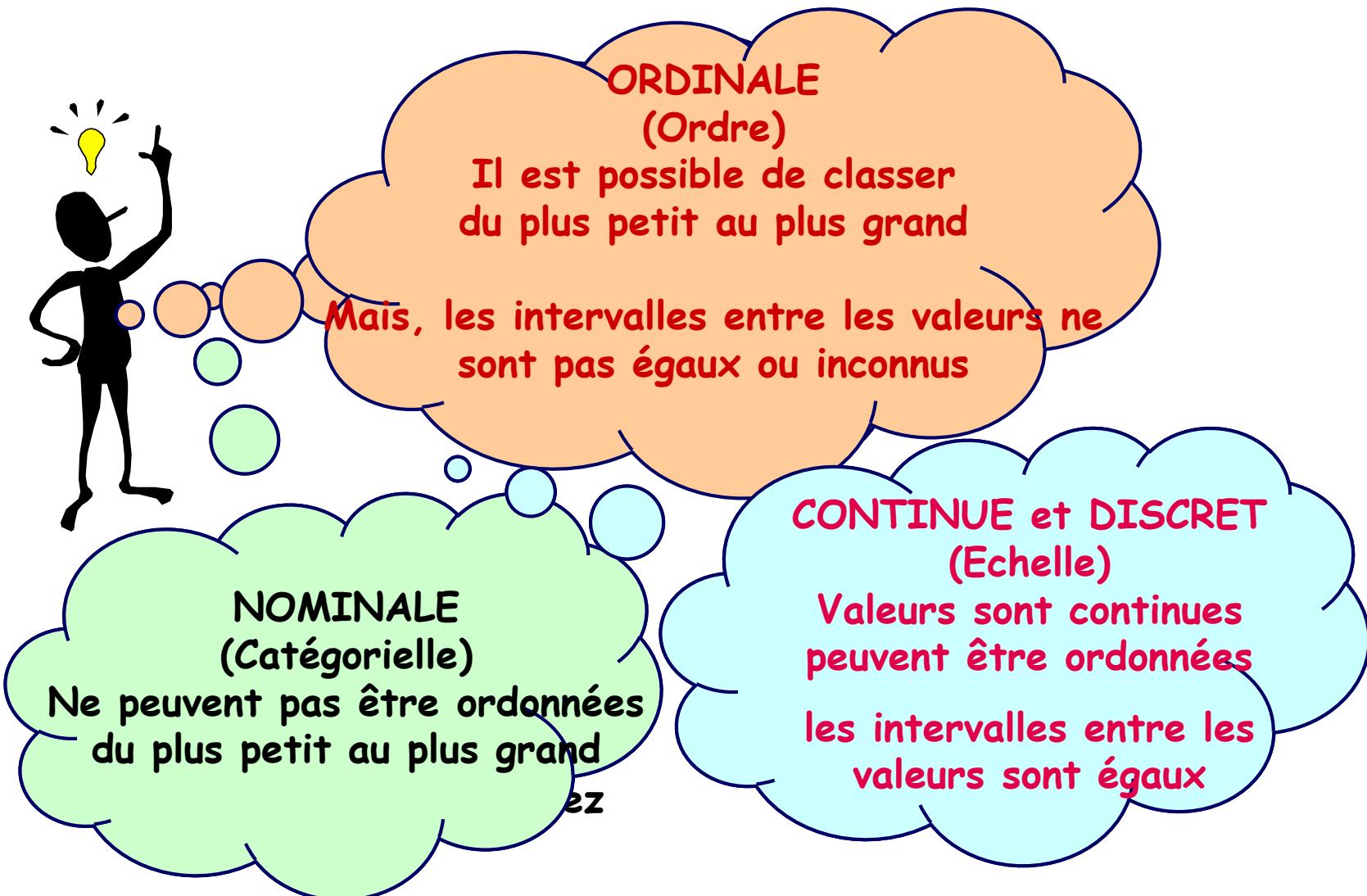
Collecte des données

La taille de l'échantillon

- Fixée en valeur absolue ou en valeur relative : fraction de sondage
- **La précision dans une enquête** dépend :
 - de la taille de l'échantillon
 - du caractère plus ou moins homogène ou hétérogène de la population parent.
- **La précision est d'autant meilleure que la taille de l'échantillon** est importante et que la population est homogène.

=> Pas de recette : pour fixer la taille d'un échantillon il est nécessaire d'avoir une idée suffisante de la précision souhaitée (risque accepté) et d'autre part du degré d'homogénéité (variabilité) de la population étudiée.

Niveaux de mesures



La collecte des données

La définition des observations

- Les observations à réaliser doivent être parfaitement définies.

ID	SEXE	DATENAIS	EDUC	CATEMP	SALDEB	SALACT	TEMPS	EXP
1	m	3-févr-1952	15	3	27 000	57 000	98	144
2	m	23-mai-1958	16	1	18 750	40 200	98	36
3	f	26-juil-1929	12	1	12 000	21 450	98	381
4	f	15-avr-1947	8	1	13 200	21 900	98	190
5	m	9-févr-1955	15	1	21 000	45 000	98	138
6	m	22-août-1958	15	1	13 500	32 100	98	67
7	m	26-avr-1956	15	1	18 750	36 000	98	114
8	f	6-mai-1966	12	1	9 750	21 900	98	0
9	f	23-janv-1946	15	1	12 750	27 900	98	115
10	f	13-févr-1946	12	1	13 500	24 000	98	244
11	f	7-févr-1950	16	1	16 500	30 300	98	143
12	m	11-janv-1966	8	1	12 000	28 350	98	26
13	m	17-juil-1960	15	1	14 250	27 750	98	34
14	f	26-févr-1949	15	1	16 800	35 100	98	137
15	m	29-août-1962	12	1	13 500	27 300	97	66

- S'il s'agit d'**observations qualitatives** i.e : résultat du classement de l'observation dans un groupe,

Nominal :

Sexe : Féminin , Masculin

Catemp : Cadre supérieur, Directeur, etc.

Échelle nominale : les codes utilisés ne servent qu'à identifier la modalité à laquelle appartient l'individu.

Dans cette échelle, il n'y a pas de relations d'ordre entre les codes.

La collecte des données en bref

La définition des observations

- S'il s'agit **d'observations ordinaires, les codes utilisés permettent :**
 - **d'identifier la modalité** à laquelle appartient l'individu ;
 - d'établir **une relation d'ordre** entre les modalités observables et par le fait même entre les individus.

Exemples :

- **Potentiel entrepreunarial** : faible ; moyen; élevé
- **Groupe d'âge** : moins de 18; de 18 à 24; 25 à 29; 30 à 34; etc
- **Niveau de scolarité** : primaire, secondaire, collégial, universitaire
- **Niveau d'appréciation** : très bonne qualité; bonne qualité; qualité moyenne

La collecte des données en bref

La définition des observations

ID	SEXE	DATENAIS	EDUC	CATEMP	SALDEB	SALACT	TEMPS	EXP
1	m	3-févr-1952	15	3	27 000	57 000	98	144
2	m	23-mai-1958	16	1	18 750	40 200	98	36
3	f	26-juil-1929	12	1	12 000	21 450	98	381
4	f	15-avr-1947	8	1	13 200	21 900	98	190
5	m	9-févr-1955	15	1	21 000	45 000	98	138
6	m	22-août-1958	15	1	13 500	32 100	98	67
7	m	26-avr-1956	15	1	18 750	36 000	98	114
8	f	6-mai-1966	12	1	9 750	21 900	98	0
9	f	23-janv-1946	15	1	12 750	27 900	98	115
10	f	13-févr-1946	12	1	13 500	24 000	98	244
11	f	7-févr-1950	16	1	16 500	30 300	98	143
12	m	11-janv-1966	8	1	12 000	28 350	98	26
13	m	17-juil-1960	15	1	14 250	27 750	98	34
14	f	26-févr-1949	15	1	16 800	35 100	98	137
15	m	29-août-1962	12	1	13 500	27 300	97	66

S'il s'agit d'**observations quantitatives** (résultat d'une mesure ou d'un comptage),

les opérations arithmétiques ont un sens

Saldeb: Salaire du début

Temps: Ancienneté de l'employé

EDUC : Nb d'années d'étude

Une variable quantitative peut être discrète ou continue :

Discrète : ne peut prendre qu'un Nb. limité de valeurs (souvent entières)

Continue : peut prendre toutes les valeurs d'un intervalle fini ou infini

En résumé : Niveaux de mesure



<i>Niveau de mesure</i>	<i>Propriété</i>		
Nominale	qualitative		
Ordinale	qualitative	Rang	
Continue/Discret Echelle (Intervalle/Ratio)	quantitative	Rang	intervalle

Exercice d'apprentissage

Précisez le type de caractère et le type d'échelle correspondant à chacune des situation suivantes :

Q1: Nb d'années d'existence de l'entreprise :

	code
Moins de 2 ans	1
2 mais moins de 5 ans	2
5 mais moins de 10 ans	3
10 ans et plus	4

Q2: Taille de l'entreprise : code

Petite	1
Moyenne	2
Grande	3
Très grande	4

Q3: Chiffre d'affaire de l'entreprise

code

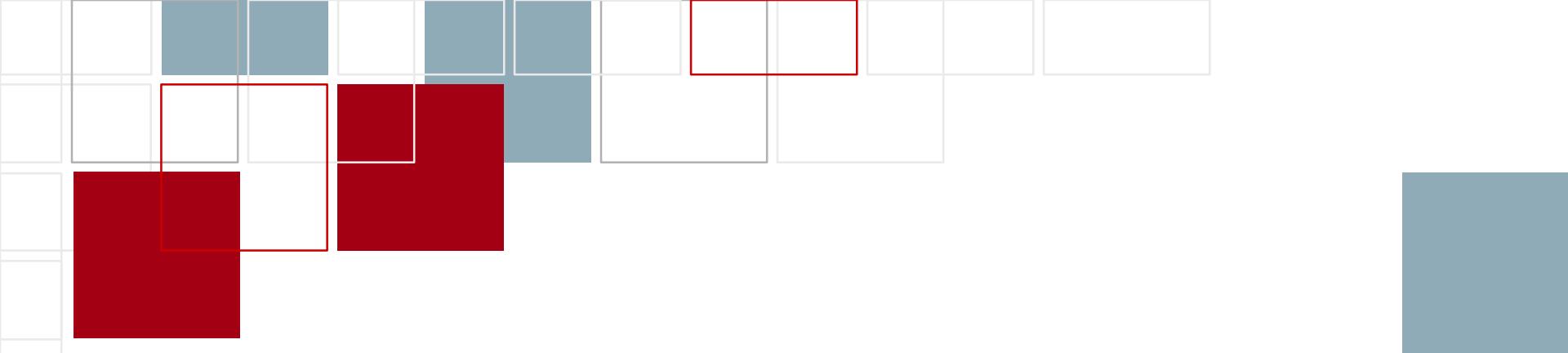
Moins de 50 000 Dh	1
50 000 Dh à 100 000 Dh	2
100 000 Dh à 500 000 Dh	3
500 000 Dh et plus	4

Q4: Nb de paintes/j au service Client

2; 4; 0; 12; 14; 8;

Q5: Opinion sur le DG choisi
Code

Pour	1
Contre	2
Ne sais pas	3



Chapitre 1 : Représentations graphiques des données

- Variables qualitatives
- Variables quantitatives discrètes
- Variables quantitatives continues

Chapitre 2 : Caractéristiques de position, de dispersion et de forme

- Variables qualitatives
- Variables quantitatives discrètes
- Variables quantitatives continues

Comment peut-on résumer les différents types de mesures ?

Qualitative

Nominale



Exemple: Couleur des yeux, sexe, etc.
Graphe: Diagramme sectoriel / Bâton
Tendance centrale : Mode

Ordinal



Exemple: niveau de satisfaction, score, etc.
Graphe: Diagramme sectoriel / Bâton
Tendance centrale : Mode + Médiane, Rang

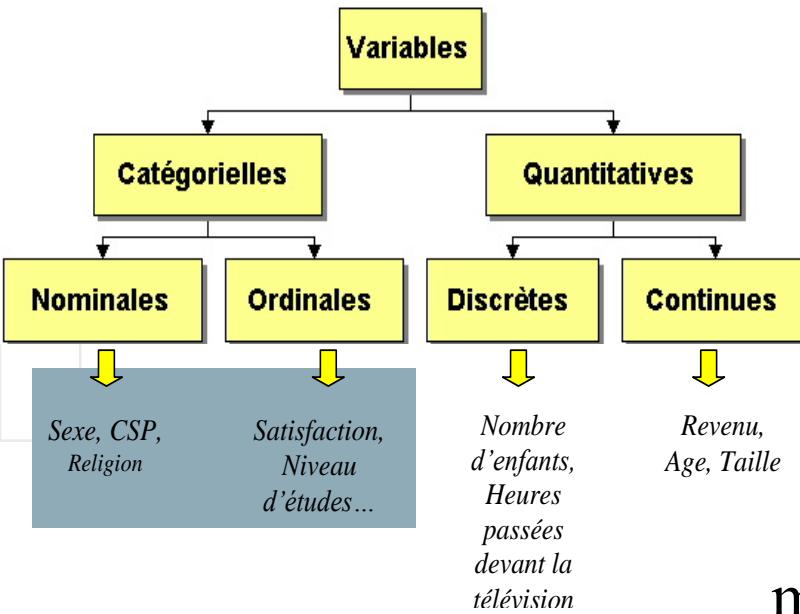
Quantitative (Echelle)

Ratio / Intervalle



Exemple: Revenu, Taille, Poids, âge, etc
Graphe: Histogramme
Tendance centrale : Mode + Médiane +
Moyenne, Rang, Ecart type

Variables Qualitatives



Soit x une variable qualitative à k modalités. L'ensemble des n individus peut être subdivisé en k groupes sur lesquels x est constante.

Si la variable x est **ordinale**, les modalités sont écrites dans l'ordre :
modalité $1 < \text{modalité } 2 < \dots < \text{modalité } k$.

Les graphiques les plus utilisés sont les suivants :

- *diagrammes circulaires* ou « en camembert »
- *diagrammes en barres*.

Dans les deux cas, les surfaces symbolisant les différentes modalités doivent être **proportionnelles** aux **effectifs**, ou aux **fréquences associées**.

Représentation graphique d'une variable qualitative

Exemple : la variable x désigne la situation familiale des individus

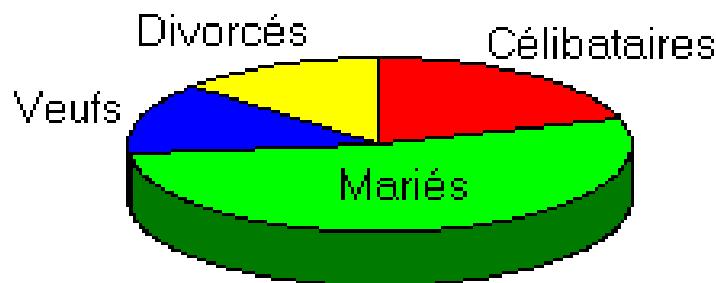


Diagramme secteuriel

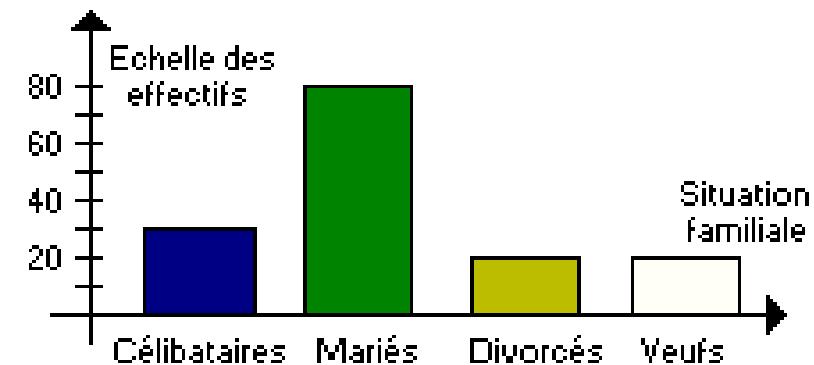
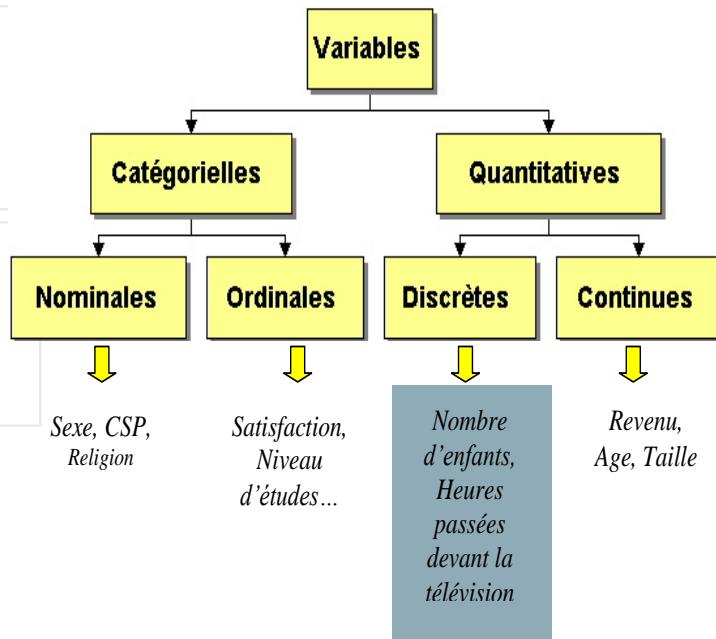


Diagramme en bâtons

Variable quantitative discrète



On rappelle qu'une variable quantitative discrète est une variable ne prenant que des valeurs entières (plus rarement décimales). **Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible.**

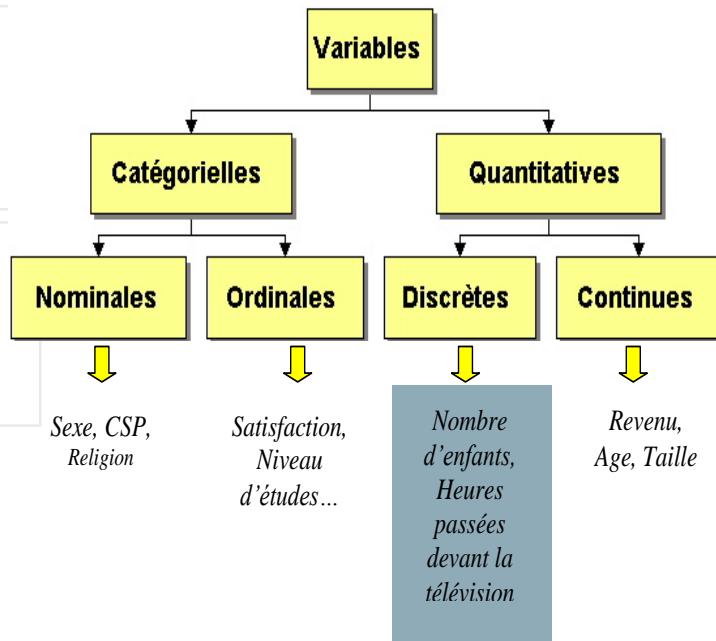
On notera dans la suite x la variable discrète et ses valeurs distinctes $\{x_1, x_2, \dots, x_p\}$. Chaque **modalité** x_i étant d'effectif (ou de **fréquence absolue**) n_i et de fréquence relative f_i , (n_i / n) $i=1, \dots, p$.

Exemple : On a noté l'âge des 48 salariés d'une entreprise, la série statistique brute est donnée ci-dessous

43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31
62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59
57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

Variable quantitative discrète

Présentation des données : Tableau statistique



xi	ni	Ni	fi(%)	Fi(%)
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,92
52	3	38	6,25	79,17
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,0

Colonne 1: ensemble des observations distinctes de x rangées par ordre croissant et **non répétées**

Colonne 2 : les **effectifs** (nombre de réplications)

Colonne 3 : Les effectifs cumulés sont définis par :

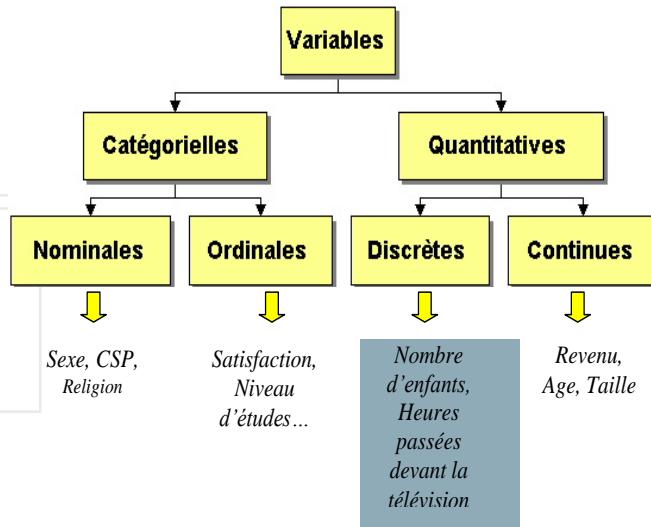
$$N_i = n_1 + \dots + n_i, \text{ pour } i \geq 1.$$

Colonne 4 : fréquences cumulées $F_i = f_i + \dots + f_1,$

↑ ↑
Effectifs **Fréquences**
cumulés cumulées

Variable quantitative discrète

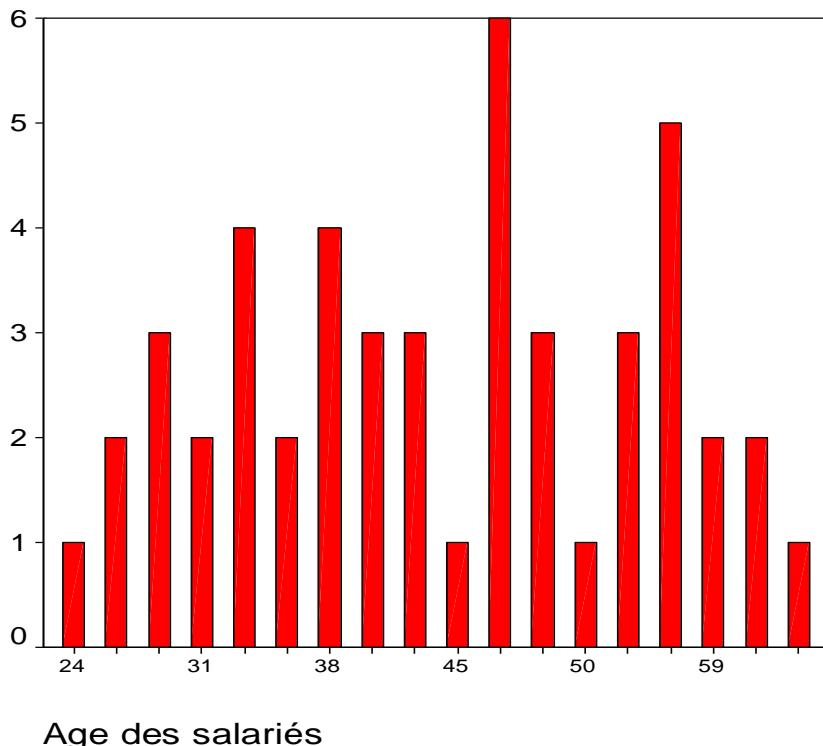
Représentation graphique (diagramme)



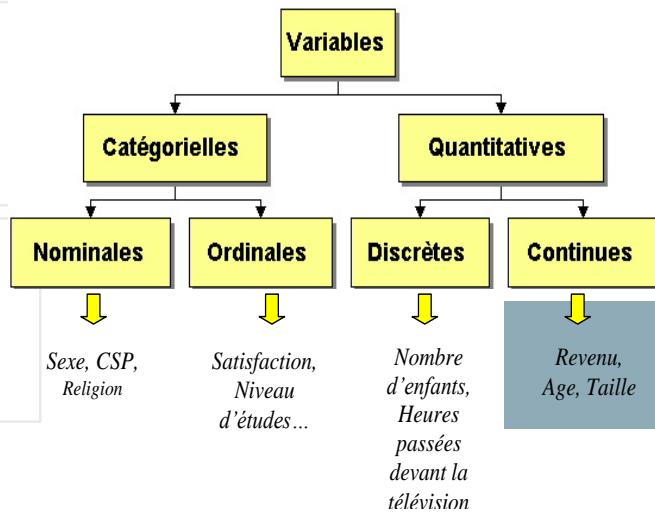
Deux graphiques populaires :

- **Diagramme en bâtons**
- **Diagramme cumulatif**

Diagramme en bâtons



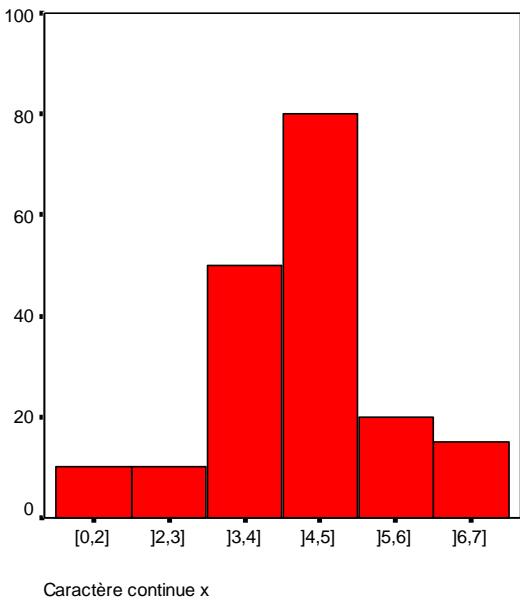
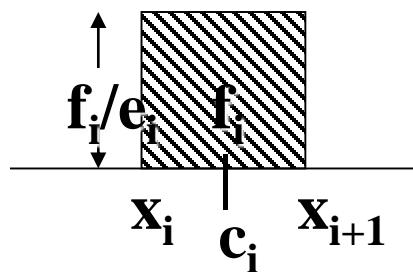
Variables quantitatives continues



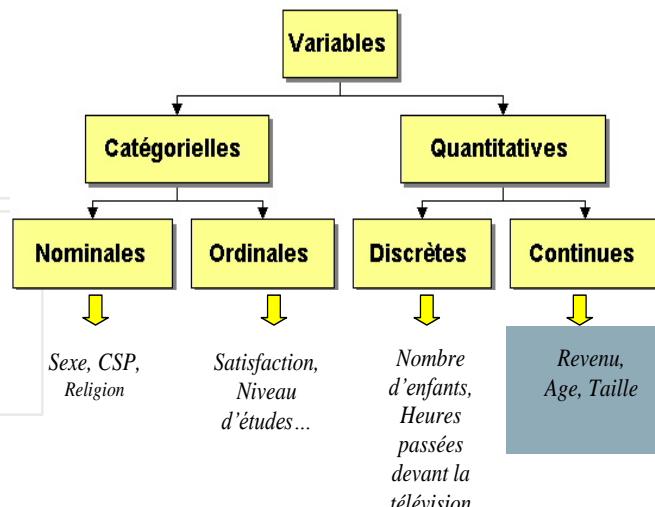
Histogramme : remplace le diagramme en bâtons pour var discrètes.

A chaque classe $[x_i, x_{i+1}]$ est associé un rectangle dont la surface est proportionnelle à la base $e_i = x_{i+1} - x_i$.

Rappelons qu'une variable quantitative est dite **continue** lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des **intervalles réels**.



Variables quantitatives continues



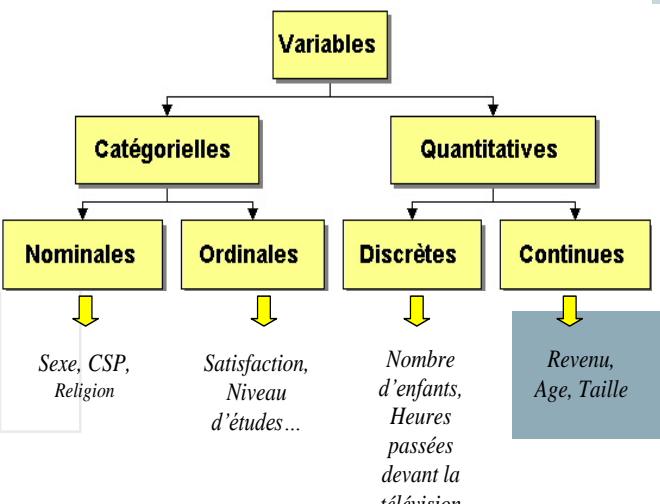
Cas d'une série chronologique ou temporelle

Une série chronologique est une série de données de nature quantitative qui ont été obtenues dans le temps à intervalles de temps réguliers.

Evolution mensuelle des délégations de crédits au cours de l'exercice 2002 (en valeur et en nombre)

Données	Janv.	Févr.	Mars	Avr	Mai	Juin	JUIL	Août	Sept	Oct	Nov	Déc	Total
Montant	23%	28%	10%	10%	6%	4%	2%	3%	4%	1%	5%	3%	100%
Nombre	3%	24%	19%	13%	11%	5%	6%	2%	5%	4%	4%	4%	100%

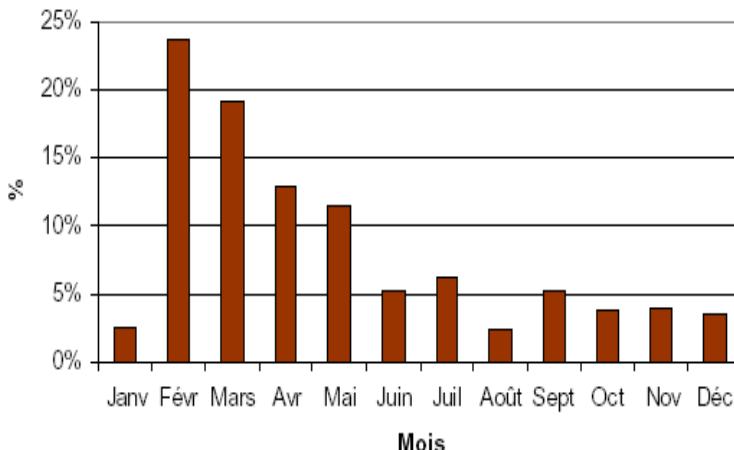
Variables quantitatives continues



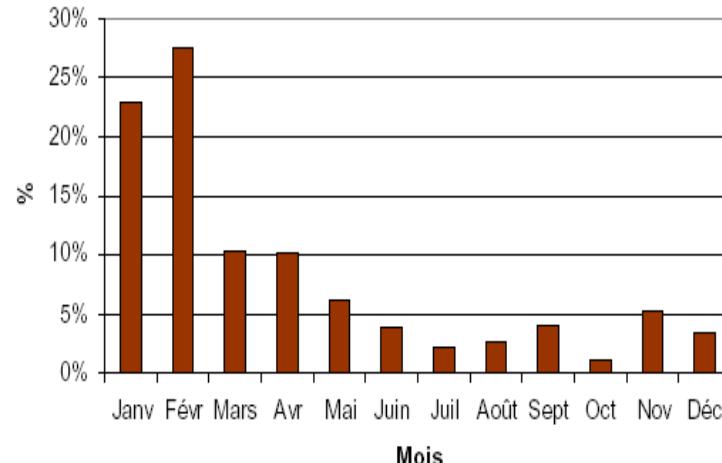
Evolution mensuelle des délégations de crédits au cours de l'exercice 2002 (en valeur et en nombre)

	Janv.	Févr.	Mars	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc	Total
Montant	23%	28%	10%	10%	6%	4%	2%	3%	4%	1%	5%	3%	100%
Nombre	3%	24%	19%	13%	11%	5%	6%	2%	5%	4%	4%	4%	100%

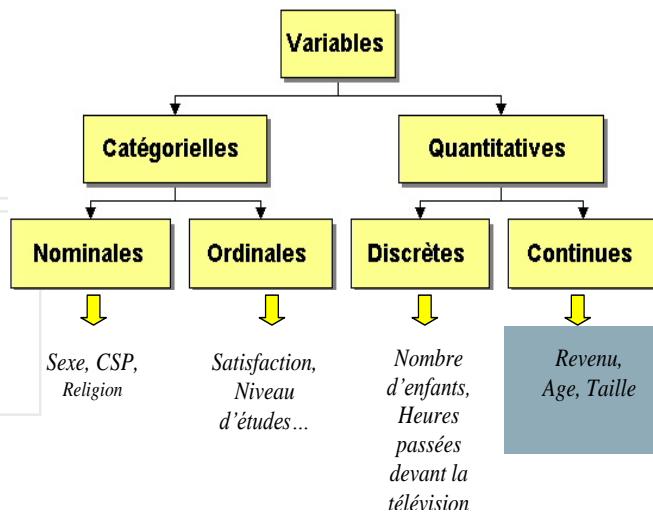
Evolution mensuelle des engagements des crédits délégués au cours de l'exercice 2002 (En nombre)



Evolution mensuelle des engagements des crédits délégués au cours de l'exercice 2002 (En Valeur)

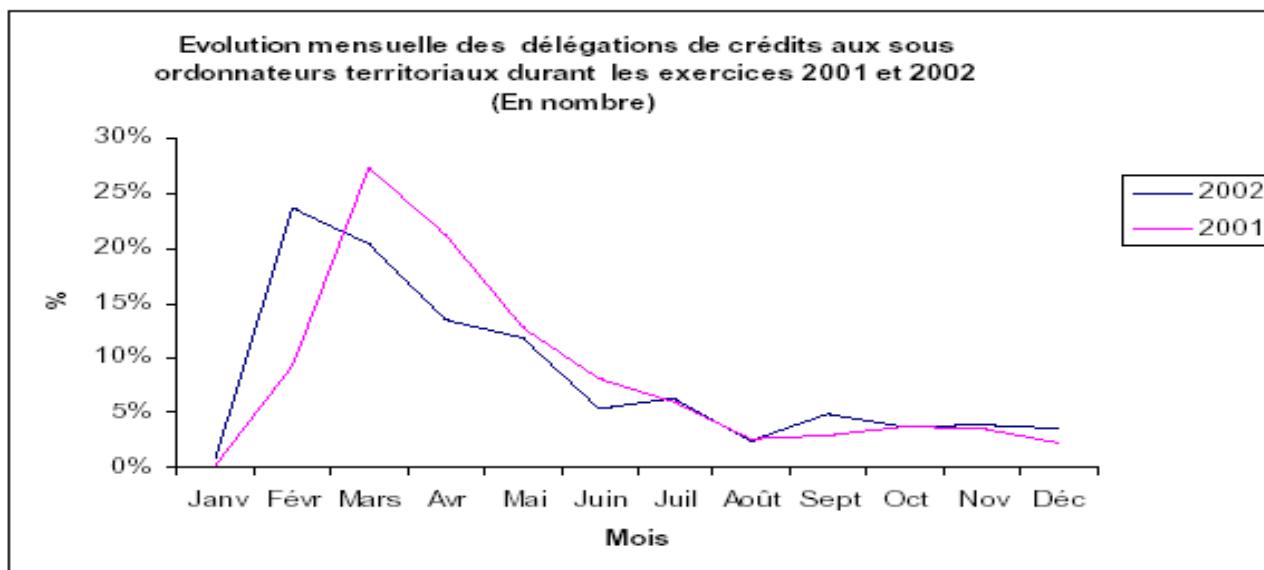


Variables quantitatives continues



Evolution mensuelle des délégations de crédits au cours de l'exercice 2002 (en valeur et en nombre)

	Janv.	Févr.	Mars	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc	Total
Montant	23%	28%	10%	10%	6%	4%	2%	3%	4%	1%	5%	3%	100%
Nombre	3%	24%	19%	13%	11%	5%	6%	2%	5%	4%	4%	4%	100%



Quels indicateurs ?

Caractéristiques de :

Tendances centrales

Dispersion

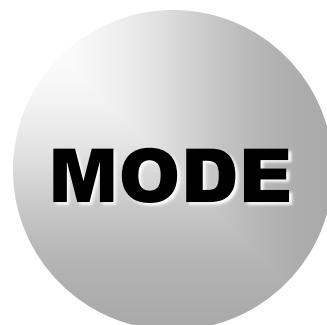
Forme

- Moyenne arithmétique
- Médiane
- Mode

- Etendu
- Ecart-type
- Indice de variabilité

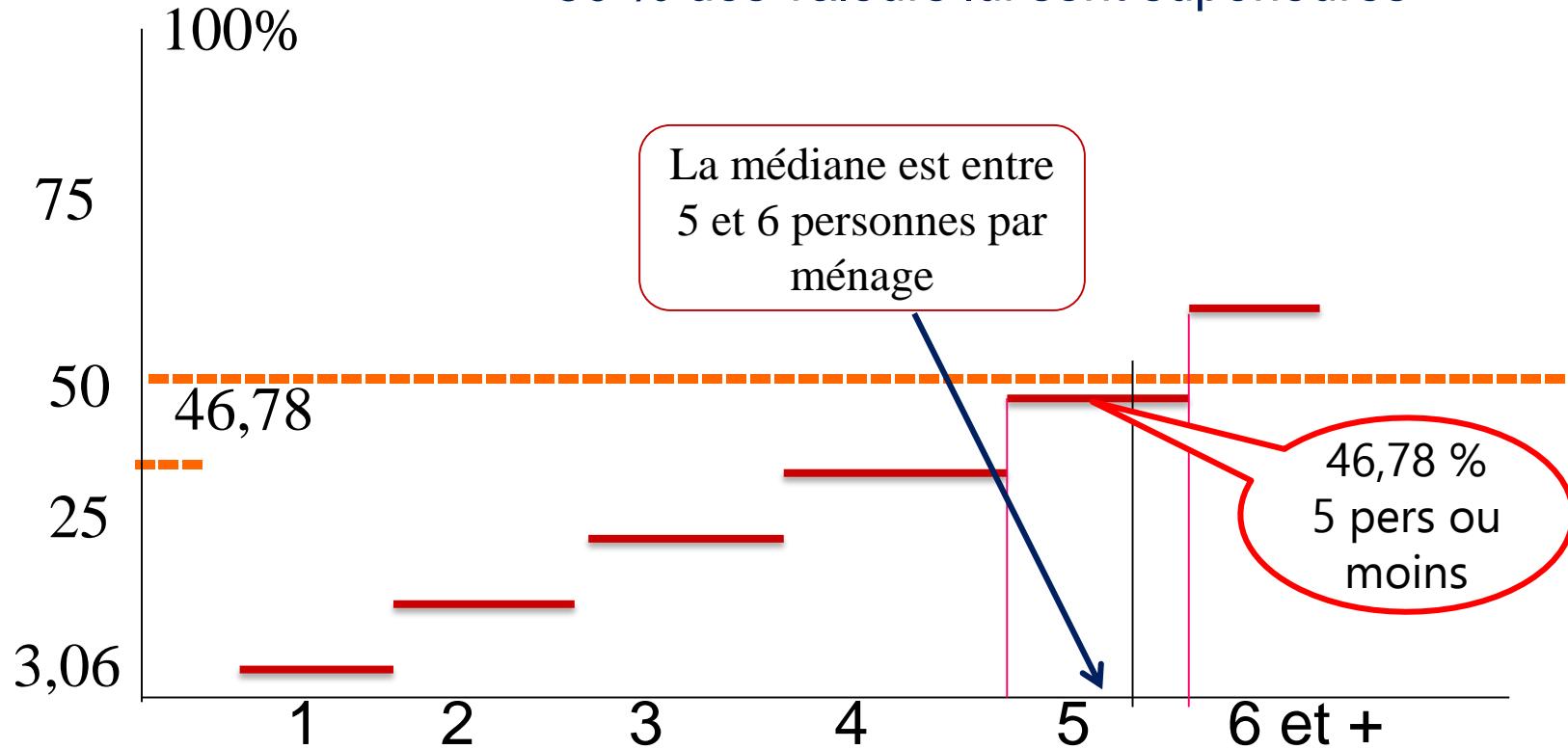
- Quantiles
- indice symétrie
- indice aplatissement

***Peut-on décrire les tendances
centrales de la même manière
pour tout type de variables ?***



La médiane

- **La médiane Q_2** : valeur du caractère qui partage les données en deux sous ensemble de même fréquence :
 - ❖ 50 % des valeurs lui sont inférieures
 - ❖ 50 % des valeurs lui sont supérieures



La médiane d'une variable discrète

Cas de valeurs distinctes

Mois	Moy T° à Essaouira
janv	3
fév	3,6
mars	6,6
avril	9,6
mai	13
juin	16
juil	17,9
août	17,7
sept	15,3
oct	11,2
nov	6,4
déc	3,7

Tri
→

Mois	Moy T° à Essaouira
janv	3
fév	3,6
déc	3,7
nov	6,4
mars	6,6
avril	9,6
oct	11,2
mai	13
sept	15,3
juin	16
août	17,7
juil	17,9

Cas pair

$$Q_2 = (9,6 + 11,2)/2 \\ = 10,4 \text{ °C}$$

1. Classer les n données dans l'ordre croissant et numérotter les valeurs ordonnées de 1 à N. Soient $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ les valeurs distinctes ordonnées.
2. Si N est impair alors $Q_2 = x_{(k)}$ avec $k = (N+1)/2$
Si N est pair alors $Q_2 = (x_{(k)} + x_{(k+1)})/2$ avec $k = N/2$

La Médiane d'une variable discrète

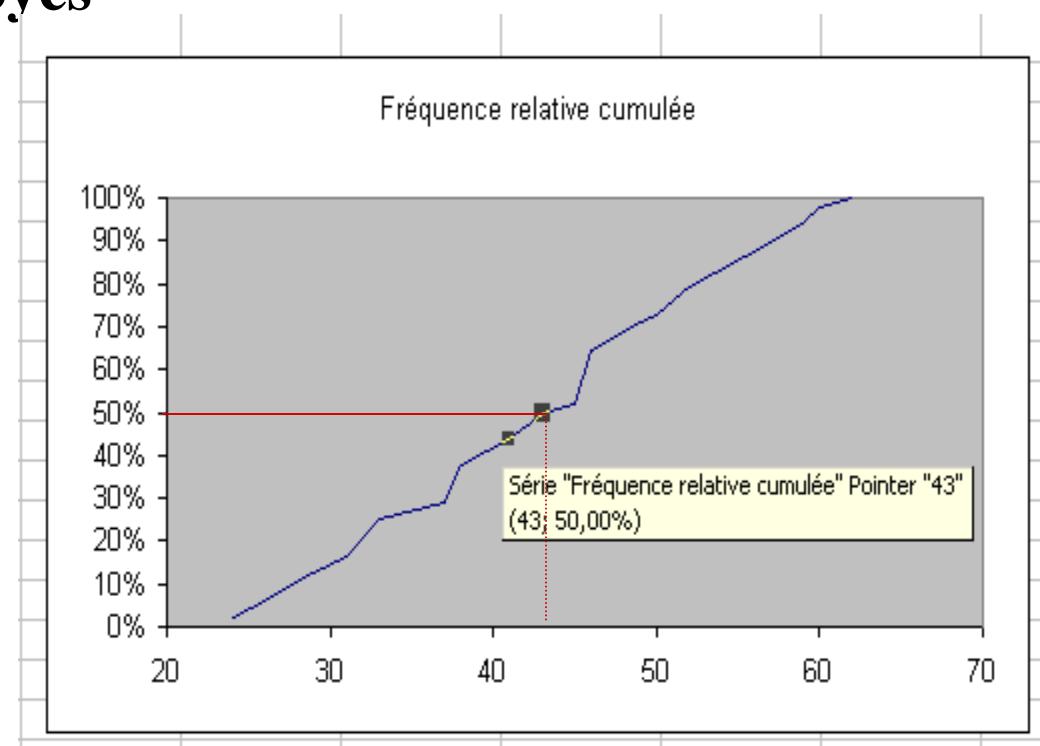
Cas de
valeurs
répétées

On utilise les **effectifs cumulés croissants** : pour les valeurs précédents Q_2 , ils sont inférieurs à $(N/2)$ et pour les valeurs suivants Q_2 , ils sont supérieurs.

Exemple de l'âge des employés

xi	ni	Ni	fi(%)	Fi(%)
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,92
52	3	38	6,25	79,17
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,0

Q₂ = 43



La Médiane

- La médiane tient compte du rang de tous les individus et non de leur valeur.
- Les valeurs exceptionnelles ne l'affectent pas.
- Elle est qualifiée d'estimateur robuste
- Valable sur caractères quantitatifs et qualitatifs ordinaux
- La médiane est la valeur centrale la plus proche de tous les individus

Médiane d'une variable continue

Cas des données groupées

Classes	Fréquence	Fi Cumulé
24 à 29,4	6	6
29,4 à 34,85	6	12
34,84 à 40,28	6	18
40,28 à 45,71	7	25
45,71 à 51,14	10	35
51,14 à 56,57	3	38
56,57 à 62	10	48

$$B_{inf} = 40,28$$

$$N/2 = 24$$

$$F = 18$$

$$F_{Me} = 7$$

$$E = 45,71 - 40,28 = 5,43$$

On effectue **une interpolation linéaire** à l'intérieur de la classe médiane afin de trouver la valeur de l'observation centrale. La formule requise pour déterminer Q_2 est la suivante :

$$Q_2 = B_{inf} + [(N/2 - F) / f_{Me}] * E$$

B_{inf} : est la borne inférieure de la classe médiane ;

F : la somme des fréquences absolues de toutes les classes précédant la classe médiane

f_{Me} : la fréquence absolue de la classe médiane

E : l'étendu de la classe médiane

L'âge médian est l'âge de la personne qui se trouve à 24^{ième} position dans le classement par ordre croissant

La classe médiane = [40,28 ; 45,14]

$$\begin{aligned} \text{L'âge médian} &= 40,28 + ((24 - 18) / 7) * 5,43 \\ &= 41,58 \text{ ans} \end{aligned}$$

La classe médiane correspond à la valeur 50% de la Fi cumulé

➤ Les quartiles Q_1 et Q_3

- Q_1 = la valeur en dessous de laquelle se trouvent 25% des observations inférieures
- Q_3 = la valeur en dessous de laquelle se trouvent 75% des observations inférieures
- Pour les valeurs précédant Q_1 , les effectifs cumulés sont inférieurs à $(N/4)$ et pour les valeurs suivants Q_1 , ils sont supérieurs.
- Pour les valeurs précédant Q_3 , les effectifs cumulés sont inférieurs à $(3 N/4)$ et pour les valeurs suivants Q_3 , ils sont supérieurs.

Médiane

	Q1	Q2	Q3	
I	I	I	I	I
0	25%	50%	75%	
Modalités (âge en mois)	Effectif ni	Effectif cumulé	Fréquenc e relative	Fréquence cumulée
3	1	1	0,1	0,1
4	1	2	0,1	0,2
7	1	3	0,1	0,3
8	3	6	0,3	0,6
9	1	7	0,1	0,7
10	1	8	0,1	0,8
12	1	9	0,1	0,9
14	1	10	0,1	1
Total	10			

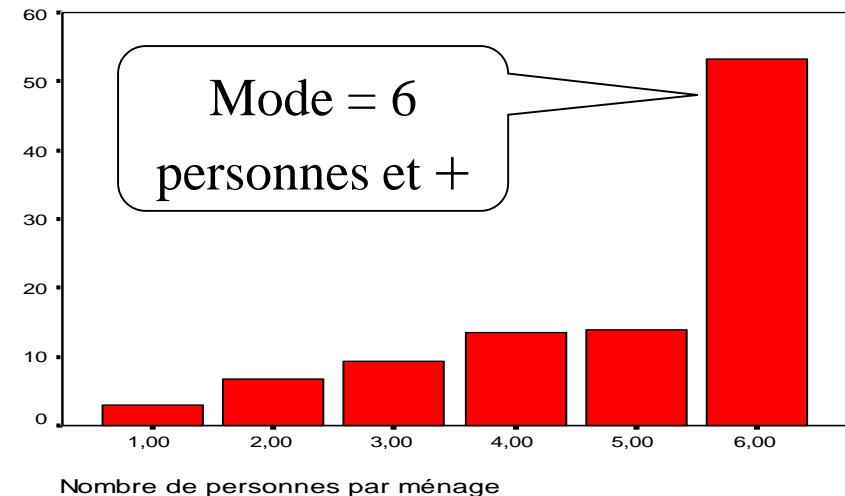
Le rang de Q_1 est : $N/4 = 10/4 = 2,5 \rightarrow$ modalité 7

Le rang de Q_3 est $3 \cdot N/4 = 3 \cdot 10/4 = 7,5 \rightarrow$ modalité 10

Le mode

- **Le mode** = la valeur du caractère la plus fréquente
- Le mode a l'avantage d'être utilisable avec les données qualitatives

Mode



Etat matrimonial	Effectifs	Pourcentage	$f_i \times 100$	$F_i \times 100$
Célibataire	158	3,1	3,1	3,1
Marié(monogame)	4200	81,9	81,9	84,9
Marié(polygame)	82	1,6	1,6	86,5
Divorcé	109	2,1	2,1	88,7
Veuf	582	11,3	11,3	100,0
Total	5131	100,0	100,0	

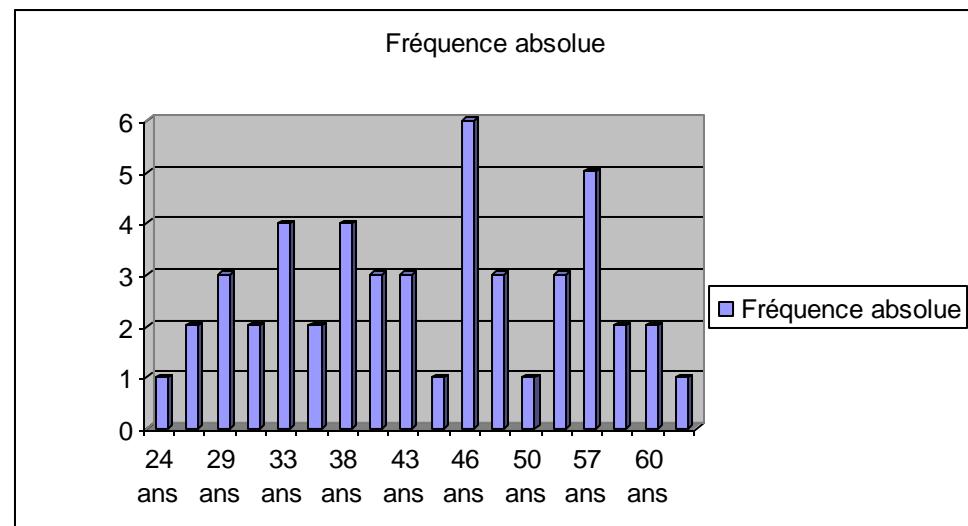
Le Mode d'une variable quantitative discrète

xi	ni	Ni	fi(%)	Fi(%)
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,92
52	3	38	6,25	79,17
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,0

46 est le mode de la variable âge

En se référant à la fréquence absolue ou effectif (colonne 2)

En se référant à la fréquence relative ou effectif (colonne 4)



Répartition des âges de 48 cadres

Moyenne arithmétique

C'est la valeur centrale la plus utilisée mais elle n'est calculable que sur des caractères quantitatifs.

S
e
n
s
i
b
l
e

caractères quantitatifs discrets :

c'est la somme des valeurs observées divisée par le nombre d'observations

caractères quantitatifs continues où les données sont groupées par classes.

On commet une légère erreur en remplaçant chacune des valeurs modalités par son centre de classe (CC)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

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

- Les valeurs extrêmes décentrent la moyenne.

Elles peuvent n'être que :

- peu significatives
- très exceptionnelles
- voir aberrantes

Il faut donc contrôler leur pertinence

Cas d'une variable continue

La moyenne

$$(18+20)/2 = 19 \text{ ans}$$

On ne considère plus les valeurs des modalités, mais les centres des classes

$$\bar{X} = \frac{1}{N} \sum_{i=1}^k n_i c_i$$

k = Nombre de classes

Classes d'âge	ni	ci	fi	Fi	ci*fi
18 - 20	10	19	0,1	0,1	1,9
20 - 22	18	21	0,18	0,28	3,78
22 - 24	23	23	0,23	0,51	5,29
24 - 26	14	25	0,14	0,65	3,5
26 - 28	10	27	0,1	0,75	2,7
28 - 30	8	29	0,08	0,83	2,32
30 - 32	4	31	0,04	0,87	1,24
32 - 34	5	33	0,05	0,92	1,65
34 - 36	1	35	0,01	0,93	0,35
36 - 38	2	37	0,02	0,95	0,74
38 - 58	5	48	0,05	1	2,4
Total	100				25,87

Autres moyennes

Attention !

La moyenne arithmétique n'est pas toujours la mieux adaptée.
C'est le cas pour les phénomènes : **multiplicatifs ; cumulatifs ; ou mettant en cause des fractions.**

Il faut utiliser une **moyenne** :

- **Géométrique** : *Quel est le taux d'accroissement annuel ?*
- **Harmonique** : *permet de calculer des moyennes de pourcentage ou des moyennes de ratios*
- **Quadratique** : *permet, de calculer des moyennes d'écart*

Mesures appropriées des tendances centrales

	Mode	Médiane	Moyenne
Nominale	✓	✗	✗
Ordinal	✓	✓	?
Continue/ Discret (Echelle)	✓	✓	✓

Comparaison des trois mesures de tendance centrale : la moyenne, la médiane et le mode

	Moyenne (\bar{x})	Médiane (Q2)	Mode (Mo_x)
Calcul	<ul style="list-style-type: none">Facile	<ul style="list-style-type: none">Difficile (il faut trier les données)	<ul style="list-style-type: none">Difficile (il faut mettre les données en classes)
Valeurs except.	<ul style="list-style-type: none">Affectent beaucoup la valeur de \bar{x}	<ul style="list-style-type: none">Affectent peu la valeur de l'Q2	<ul style="list-style-type: none">Affectent peu ou pas la valeur de Mo_x
Intérêt principal	<ul style="list-style-type: none">Bon estimateur de tendance centrale si distribution sym.\bar{x} est plus efficace que Q2	<ul style="list-style-type: none">Plus précise que Mo_xMoins affectée que \bar{x} par les valeurs extrêmes	<ul style="list-style-type: none">Pour décrire une distribution plurimodalePeut être calculé pour variables circulaireset pour var. qualitatives

Mesures de dispersion

**Votre score est de 55% dans un test.
Quelle est votre performance si le
score moyen est de 50%?**

Mieux que la moyenne?

Oui, mais de combien?

**Pour le savoir, vous avez besoin de
connaître
l'étendue/Variabilité/dispersion des
données**

Peut-on décrire les mesures de dispersion de la même manière pour tout type de données ?

Variance

SD

MIN
/MAX

IQR

Mesures de dispersion

- Dans la plupart du temps, les mesures de tendance centrale ne peuvent à elles seules décrire et résumer convenablement un ensemble de données.
- Exemple 1** Dans le service d'urgence d'un hôpital on note à chaque intervalle de temps d'une heure le nombre d'arrivées de malades ou de blessées (l'observation a durée 12 heures). Les résultats sont données dans le tableau suivant
- Le service d'urgence traite en moyenne 6 patients par heure.
- La dispersion du nombre d'arrivés de cas urgents est en général très grande, il se peut très bien que, durant une certaine heure, il n'y ait qu'un seul arrivé ou aucun et que durant l'heure suivante il y en ait 12 ou 17. c'est le cas observé dans cet exemple.
- Pour éviter que le service soit trop souvent débordé. On doit l'organiser de telle sorte qu'il soit en mesure de traiter, par moments beaucoup plus que 6 patients par heure.
- La demande moyenne d'un service est un indice inadéquat des ressources nécessaires à sa prestation

Intervalle de temps	Nombre d'arrivées
1	5
2	2
3	4
4	3
5	0
6	12
7	10
8	17
9	11
10	1
11	2
12	5
Total	72

Mesures de dispersion

■ Exemple 2

	Moyenne	Médiane	Ecart-type
Ensemble 1 : 20, 20, 20	20	20	0
Ensemble 2 : 10, 20, 30	20	20	8.16
Ensemble 3 : 1, 20, 39	20	20	15.51

- Dans les trois cas, la moyenne est égale à 20, ainsi que la valeur de la médiane.
- On ne saurait pour autant conclure que les trois ensemble sont identiques.
- → la variabilité des données est plus grande dans l'ensemble 3 que dans les deux ensembles 2 et 1

Mesures de dispersion

- **L'étendu :** c'est la différence entre les valeurs extrêmes du caractère x observé : $E = \max(x_i) - \min(x_i)$
- **La variance, l'écart type**
- **L'écart interquartiles**
- **Le coefficient de variation**
- **Variance et écart type**
 - La variance d'une distribution de fréquence est la moyenne arithmétique des carrés des écarts à la moyenne.
 - Sert à Caractériser de façon globale l'écart plus ou moins important de l'ensemble des valeurs de la distribution par rapport à la valeur moyenne.
 - La variance permet de comparer la dispersion de deux ou plusieurs distributions d'une même variable.
 - Dans le cas d'une variable continue groupée en classes on utilise les centres de classes à la place des x_i .

$$S^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N} \sum_{i=1}^k n_i (x_i - \bar{x})^2$$

$$S^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2 \quad f_i = \frac{n_i}{N}$$

$$S = \left(\sum_{i=1}^k f_i (x_i - \bar{x})^2 \right)^{1/2}$$

Exemple 1. Comparaison du nombre d'arrivés/heure de cas urgent dans deux hôpitaux différents

$$\bar{x}_{h1} = 5, \quad \bar{x}_{h2} = 5$$

$$s_{h1} = 1.6, \quad s_{h2} = 3.5$$

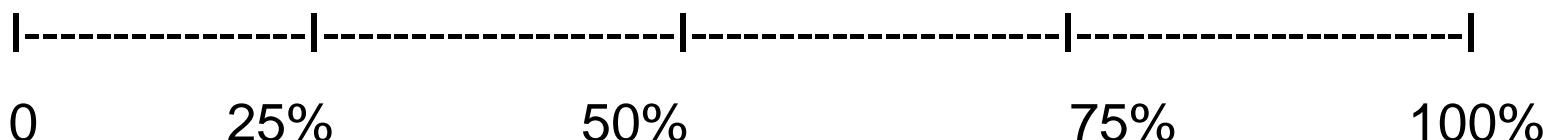
La variabilité des arrivées dans h2 est plus grande que celle des arrivés dans h1.

$$S^2 = \sum_{i=1}^k f_i (c_i - \bar{x})^2$$

L'écart interquartile

- L'écart interquartile = comprend 50% des observations, celles qui sont les plus centrales.

← Écart interquartile →



- L'écart interquartile = l'espace compris entre les quartiles 1 et 3

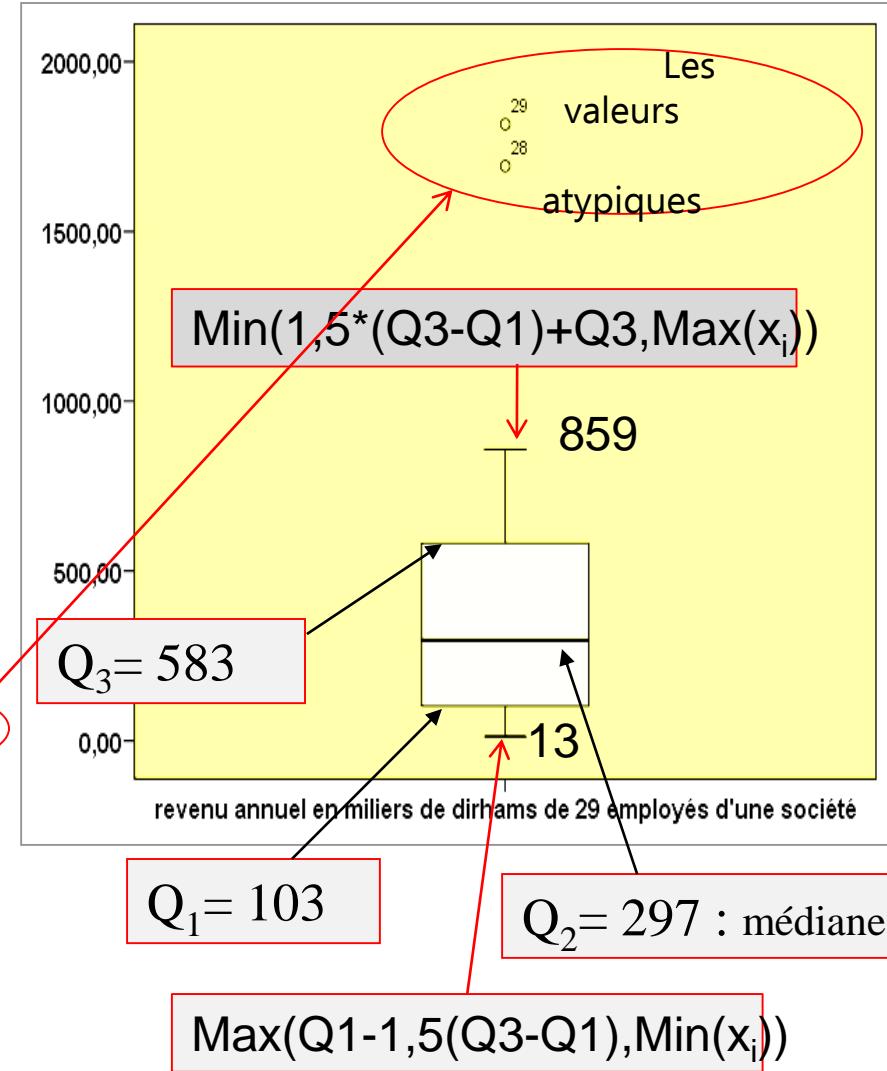
$$EQ = Q3 - Q1$$

- **EQ** est bien une mesure de dispersion, puisque plus les observations sont concentrées, plus Q_1 et Q_3 sont rapprochés et donc plus **EQ** est petite.
- **EQ** est moins utilisée que l'écart type.
- Est la mesure la plus appropriée pour des distributions fortement dissymétriques.

Boite à moustaches

- La boite à moustache ou box-plot est un résumé graphique d'une distribution.
- Le corps de la boite est formé par le premier et troisième quartile et coupé par le deuxième quartile (médiane) plus deux autres valeurs qui sont $\text{Min}(1,5*(Q3-Q1)+Q3, \text{Max}(x_i))$ et $\text{Max}(Q1-1,5(Q3-Q1), \text{Min}(x_i))$.

13	45	222	335	492	711
17	94	248	375	583	859
19	103	290	387	609	1693
31	104	295	444	618	1816
42	217	297	463	700	



Boîte à moustaches

- On repère sur la boîte à moustaches d'une variable:
- L'échelle des valeurs de la variable, située sur l'axe vertical.
- La valeur du 1er quartile Q1 (25% des effectifs), correspondant au trait inférieur de la boîte.
- La valeur du 2ème quartile Q2 (50% des effectifs), représentée par un trait horizontal à l'intérieur de la boîte.
- La valeur du 3ème quartile Q3 (75% des effectifs), correspondant au trait supérieur de la boîte.
- Les 2 moustaches, délimitent les valeurs dites *adjacentes qui sont* déterminées à partir de l'écart interquartile (Q3-Q1).
- Les valeurs dites extrêmes, atypiques, exceptionnelles, (*outliers*) situées *au-delà des valeurs adjacentes* sont individualisées. Elles sont représentées par des marqueurs (o, ou *, etc.).

Coefficient de variation

- Les mesures de dispersion considérées dans ce qui précède sont des mesures de dispersion absolue.
- Donc elles ne permettent pas de comparer la dispersion de deux ou plusieurs distributions d'une même variable mais de tendances centrales différentes
- **Exemple 1** : Considérons les deux distributions suivantes
- L'écart-type est presque le même dans les deux distributions
- **Moyenne (Médecins) ≠ Moyenne (infirmiers)**
- Nous nous pouvons pas comparer ces deux distributions en terme de dispersion.

Médecins xi	Fréquence ni	ni*xi	Infirmiers xi	Fréquence ni	ni*xi
1	4	4	2	3	6
2	4	8	3	3	9
3	5	15	4	6	24
4	2	8	5	3	15
Total	15	35	Total	15	54

Moyenne 2,33

Distribution du nombre de **médecins** dans 15 dispensaires d'une Wilaya

Moyenne 3,6

Ecart-type 1,055

Distribution du nombre de **infirmiers** dans 15 dispensaires d'une Wilaya

- Rien à conclure en comparant seulement les mesures de dispersion absolue.
- Afin d'effectuer des comparaisons, nous avons besoin d'une mesure du degré de dispersion relative au sein de la distribution étudiée.

Coefficient de variation

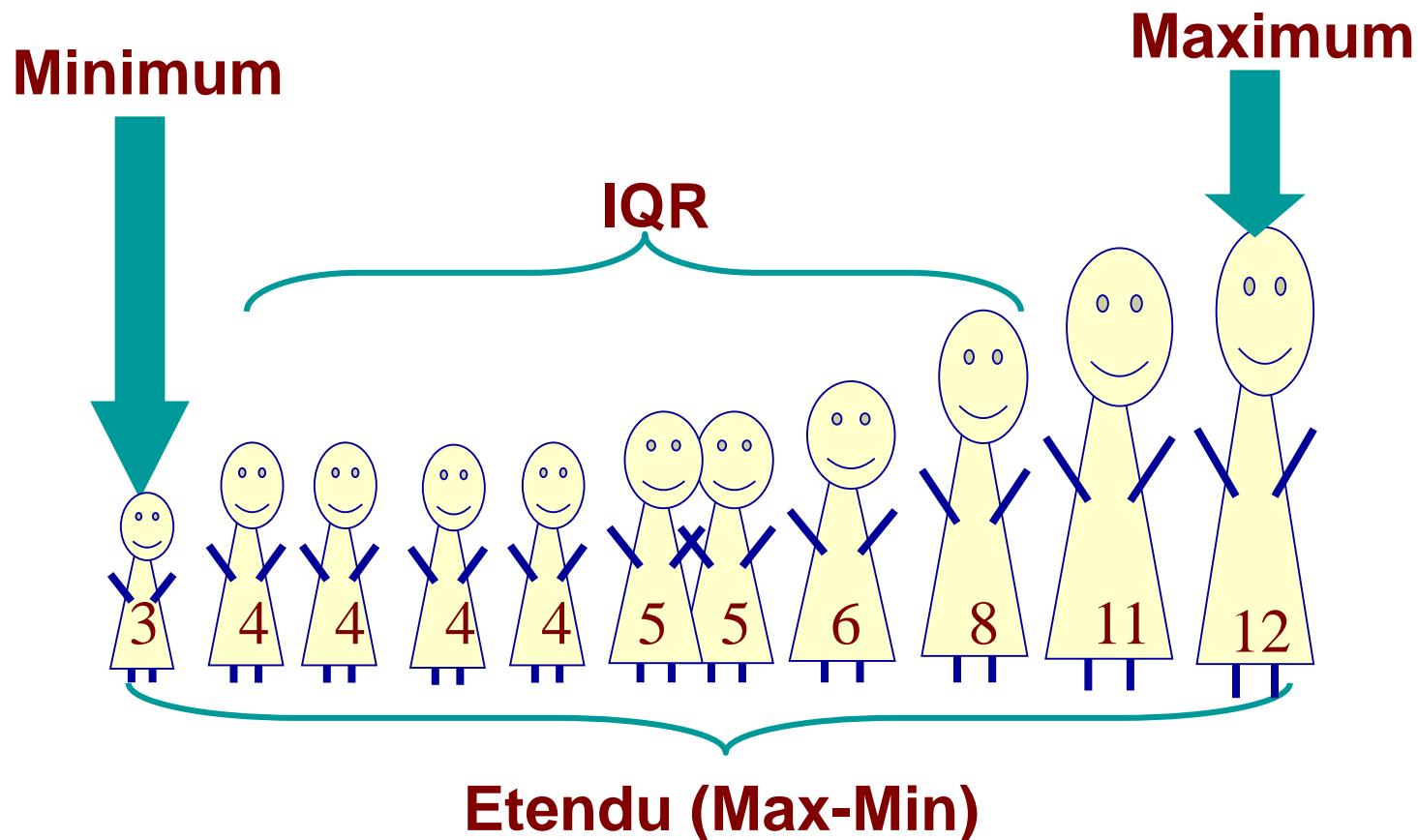
- La dispersion relative la plus utilisée est le Coefficient de variation qui correspond à l'écart type exprimé en pourcentage de la moyenne.
- C'est une mesure sans unité, donc plus pratique pour comparer deux distributions.
- Reprenons l'exemple précédent:
- La dispersion relative de la distribution des médecins est beaucoup plus grande que celle de la distribution des infirmiers.
- Le groupe des infirmiers est plus homogène que le groupe des médecins quand à leurs répartition dans les dispensaires.

$$CV = \frac{S}{\bar{X}} (100)$$

$$CV(\text{Médecins}) = 1,046/2,33 = 0,45$$

$$CV(\text{Infirmiers}) = 1,055/3,6 = 0,29$$

Mesures of Dispersion



Standard déviation: en moyenne, de combien chaque score diffère de la moyenne ?

Mesures de dispersion appropriées

	Min/Max	Variance	IQR	Standard Deviation
Nominale	✗	✗	✗	✗
Ordinal	✓	✓	✓	✗
Echelle	✓	✓	✓	✓

Mesures de dispersion

Etendu : c'est la différence entre les valeurs extrêmes du caractère

$$x : e = \max(x_i) - \min(x_i)$$

Ecart interquartile : $Iq = Q_3 - Q_1$

Ecart moyen : $e_m = \frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x}) = \sum_{i=1}^p f_i (x_i - \bar{x})$

Variance : $\sigma^2 = \frac{1}{N} \sum_{i=1}^p n_i (x_i - \bar{x})^2 = \sum_{i=1}^p f_i (x_i - \bar{x})^2$

Ecart - type : σ

Le coefficient de variation : $CV = \frac{\sigma}{\bar{x}}$

Coefficient de variation

- **Exemple 2 :**

Considérons la répartition des revenus de deux groupes d'individus (hommes et femmes) suivant :

- Bien que l'écart type des revenus des hommes (400 DH) soit inférieur à celui des revenus des femmes (800), on constate par le calcul des **CV**, que la dispersion relative des revenus des hommes est supérieure à celle des revenus des femmes.
- Autrement dit, le groupe des femmes est légèrement plus homogène que le groupe des hommes quand aux revenus annuels observés.

	Hommes	Femmes
S	400	800
\bar{x}	10000	22000
CV	4%	3,64%

L'écart type est 4% de la moyenne

L'écart type est 3,64% de la moyenne

En guise de résumé

Type de variables	NOMINALE	ORDINALE	ECHELLE
Définition	Catégories Non ordonnées	Catégories ordonnées	Valeurs Numériques
Exemples	CSP, genre, statut marital	Niveau de Satisfaction, Tranches d'âge	Revenu, Poids, âge, Taille
Mesures de Tendance Centrale	Mode	Mode Médiane	Mode Médiane Moyenne
Mesures de Dispersion		Min/Max/E EQ	Min/Max/E Variance EQ
Graphe	Secteur (Bâton)	Bâton (Secteur)	Histogramme (Bâton)
Procédures	Fréquences	Fréquences	Fréquences, Descriptives

Mesures de forme

■ Coefficient d'asymétrie (SKWENESS)

❖ Si AS = 0 , la distribution est symétrique

❖ Si AS > 0 , la distribution présente une asymétrie à gauche

❖ Si AS < 0 , la distribution présente une asymétrie à droite

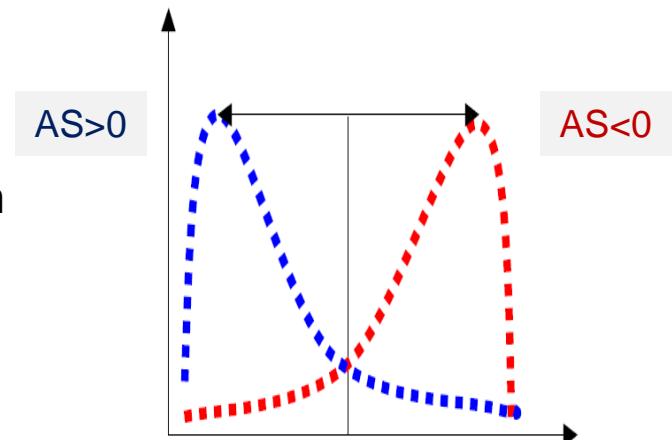
■ Coefficient d'aplatissement : KURTOSIS

➤ Si AP = 3, la distribution a un coefficient d'aplatissement similaire à la distribution normale : **Distribution mésocurtique**

➤ Si AP > 3, la distribution sera plus tassée que la distribution normale avec des queues épaisses: **Distribution Leptocurtique**

➤ Si AP < 3, la distribution présente des queues plus fines que celle de la loi normale : **Distribution Platicurtique**

$$AS = \frac{m_3}{s^3} \quad m_3 = \frac{1}{N} \sum_{i=1}^k n_i (x_i - \bar{x})^3$$



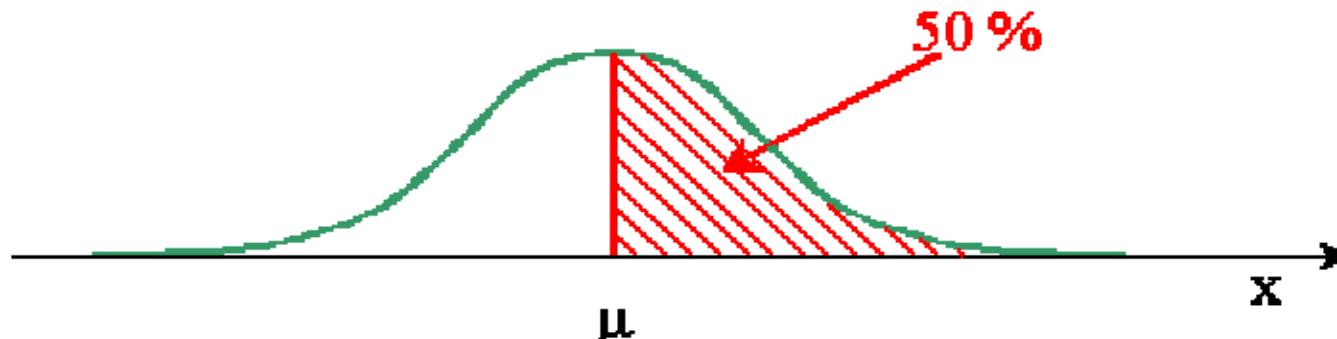
$$AP = \frac{m_4}{s^4} \quad m_4 = \frac{1}{N} \sum_{i=1}^k n_i (x_i - \bar{x})^4$$

Distribution normale

- On constate que les distributions de données continues présentent souvent une forme relativement régulière (en forme de cloche, atteignant progressivement le maximum avant de diminuer graduellement) qu'on appelle distribution normale ou distribution de Gauss.
- La distribution normale est une distribution très fréquente dans les phénomènes naturels → □ âge, hauteur, poids, erreurs aléatoires ...
- Une distribution est normale lorsque la majorité des sujets sont regroupés de façon symétrique autour de la moyenne.
- L'importance de cette distribution en statistique lui donne le nom de « distribution de référence ».
- Nous avons besoin de connaître certaines de ses propriétés.

Distribution normale

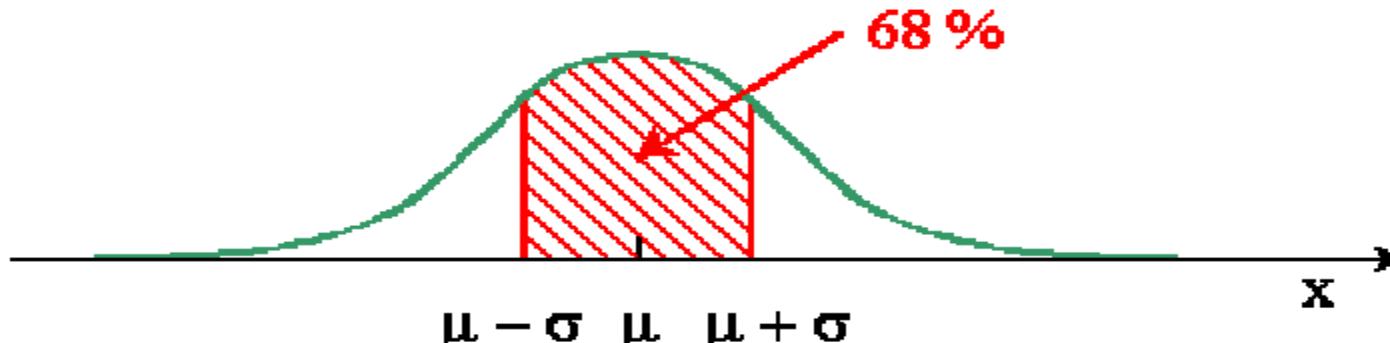
Elle est unimodale et symétrique : 50% des individus sont en dessous de la moyenne et 50% au-dessus.



La moyenne (ici appelée μ) est égale au mode et à la médiane.

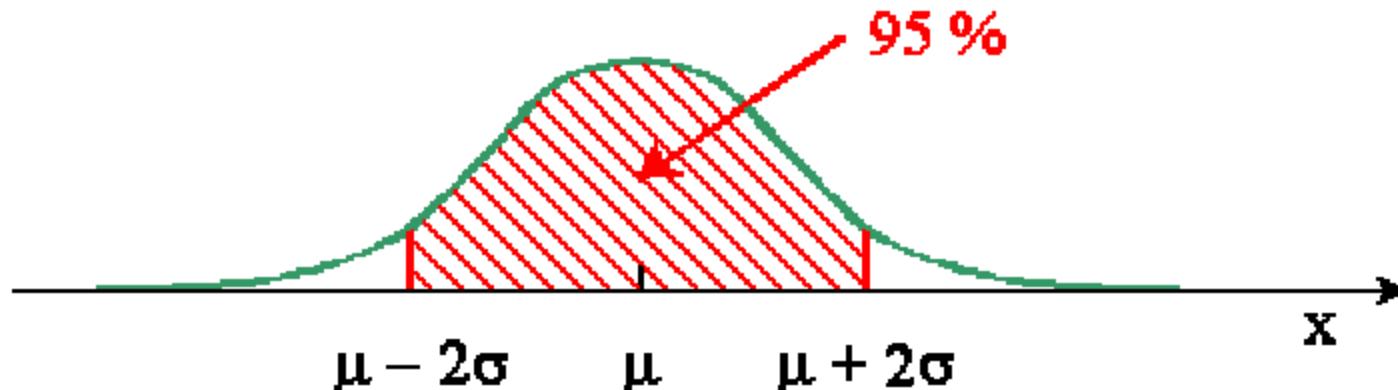
- Dans des distributions normales l'écart-type se révèle une mesure de dispersion très

On retrouve 68.26% de la population entre ± 1 écart-type (ici appelé s) autour de la moyenne

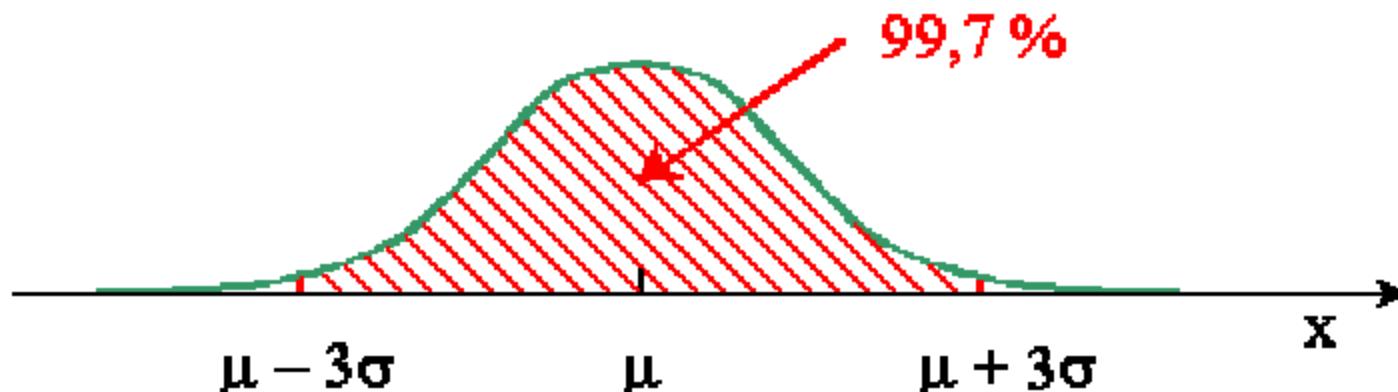


Distribution normale

On retrouve 95.44% de la population entre ± 2 écarts types autour de la moyenne.



On retrouve 99.74% de la population entre ± 3 écarts types autour de la moyenne



Vérification de la normalité d'une variable

- Distribution en forme de cloche?
- La distribution symétrique? (Histogramme)
- Vérifier si MOYENNE \approx MEDIANE \approx MODE
- Vérifier si Skewness AS \approx 0
- Vérifier si Kurtosis AP \approx 3
- Vérifier si il y a des Outliers (Explore - Box-plots)
- Diagrammes PP et QQ

Analyse → Statistiques Descriptives → Explorer

Scores standardisés

- **Le score Z** (ou score standard) : est une technique statistique qui consiste à convertir un score individuel en un score standardisé, encore appelé score centré et réduit ou score Z.
- **Le score Z** permet de fournir une indication précise de la position du score de l'individu au sein de la distribution.
- **Le score Z** indique de combien en écart-type s'écarte une observation de sa moyenne.
- Pour comparer deux distributions obtenues sur des échelles d'intervalle d'un même échantillon, on transforme les données de chaque distribution en scores centrés réduits.
- Cette transformation consiste essentiellement à exprimer les données dans un système de mesure standard, correspondant à la courbe normale centrée réduite, symbolisé par Z

$$Z = (X - \mu) / s$$

Pour une distribution normale centrée réduite on a:

- Mode = médiane = moyenne = 0
- l'écart-type vaut toujours 1 (s =1)

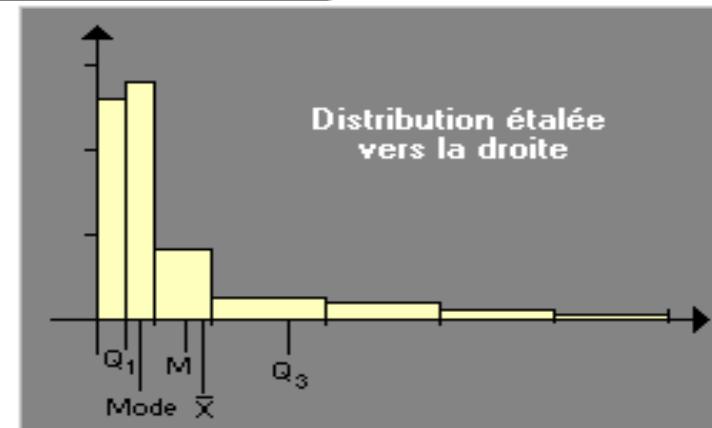
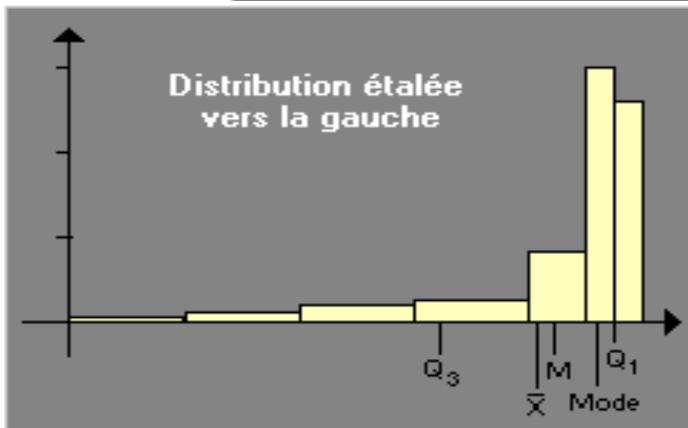
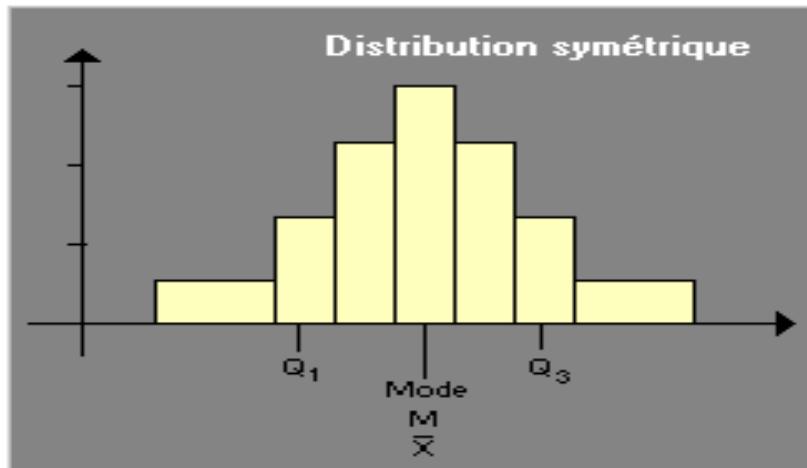
Mesures de forme

- Quantiles
- indice symétrie
- indice aplatissement

Cf. Fiche 6

Mesures de forme

Exemple:



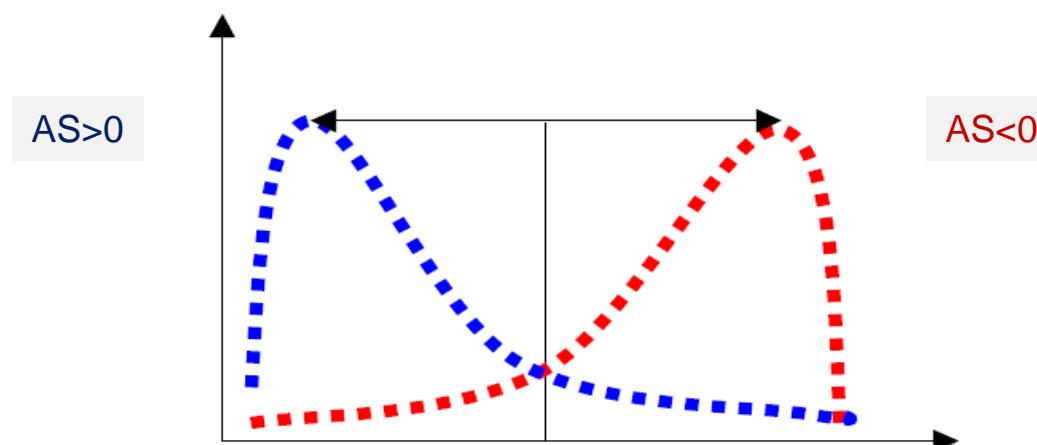
Coefficient d'asymétrie : SKWENESS

$$S = \frac{\mu_3}{\sigma^3}$$

$$\mu_3 = E[x - E(x)]^3$$

Avec μ_3 correspondant au moment centré d'ordre 3
 σ , correspondant à l'écart type

- Si $S=0$, la distribution est symétrique comme la loi normale
- Si $S>0$, la distribution penche à droite
- Si $S<0$ la distribution penche à gauche



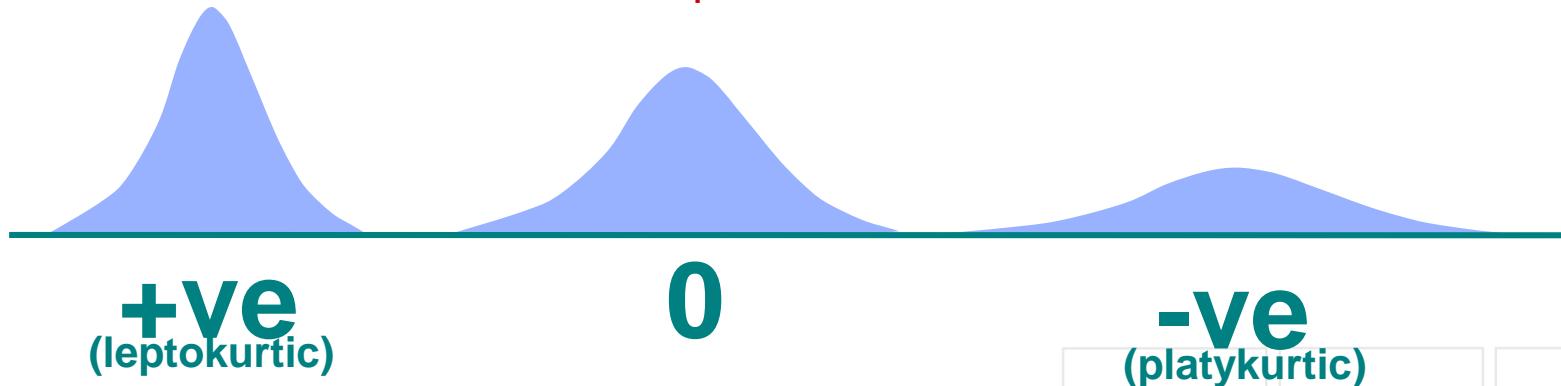
Coefficient d'aplatissement : KURTOSIS

$$K = \frac{\mu_4}{\sigma^4}$$

$$K = \frac{E[x - E(x)]^4}{(\sigma^2)^2}$$

Avec μ_4 correspondant au moment centré d'ordre 3
 σ , correspondant à l'écart type

- Si $K = 3$, la distribution a un coefficient d'aplatissement similaire à la distribution normale : **Distribution mésocurtique**
- Si $K > 3$, la distribution sera plus tassée que la distribution normale avec des queues épaisses: **Distribution Leptocurtique**
- Si $K < 3$, la distribution présente des queues plus fines que celle de la loi normale : **Distribution Platicurtique**





Chapitre 3 : Liaison entre variables

- **Liaisons entre deux variables qualitatives**

- Coefficient χ^2

- Représentation graphique entre deux variables qualitatives

- **Liaisons entre deux variables quantitatives**

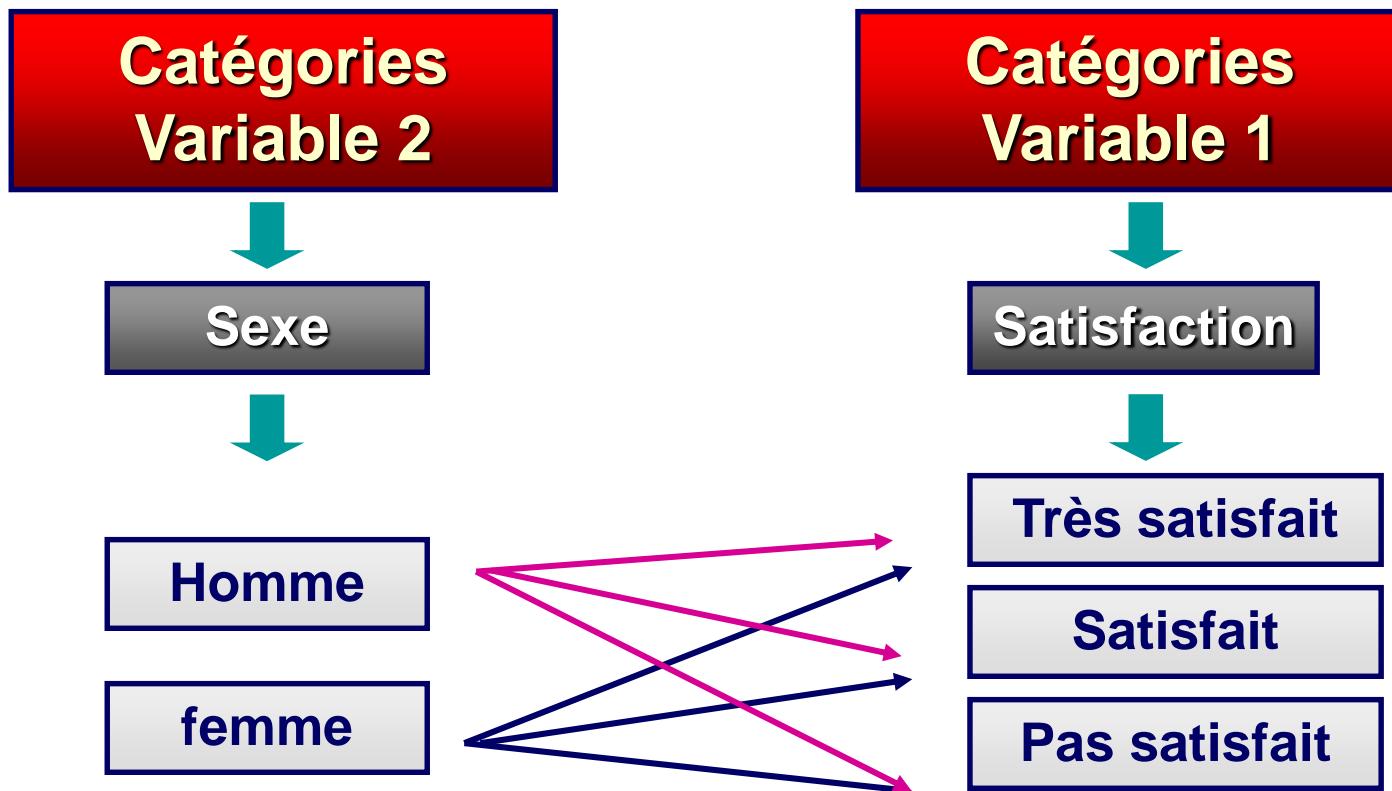
- Représentation graphique

- Coefficient de corrélation

- **Liaisons entre une variables qualitative et quantitative**

- Représentation graphique quantitative contre qualitative

Exploration de la liaison entre deux variables qualitatives (catégorielles)



Liaisons entre deux variables qualitatives

Soit x et y deux caractères qualitatifs. Les modalités de x sont notées $x_1, \dots, x_i, \dots, x_K$ et celles de y sont $y_1, \dots, y_j, \dots, y_L$.

Tableau de contingence : il est défini comme suit : pour chaque modalité x_i de x et y_j de y , n_{ij} est le nombre d'individus pour lesquels x vaut x_i et y vaut y_j . Les effectifs $n_{i\cdot}$ et $n_{\cdot j}$ désignent les totaux lignes et colonnes respectivement de ces individus:

$$n_{i\cdot} = \sum_j n_{ij}, \quad n_{\cdot j} = \sum_i n_{ij} \quad \text{avec} \quad n = \sum_{ij} n_{ij} = \sum_i n_{i\cdot} = \sum_j n_{\cdot j}.$$

x	$y_1 \dots y_j \dots y_L$	Total
x_1	\vdots	$n_{i\cdot}$
x_i	n_{ij}	
x_K	\vdots	
Total	$n_{\cdot j}$	n

Coefficient χ^2 pour deux variables qualitatives

A partir d'un tableau de contingence, on peut calculer un **coefficient χ^2** (« chi-deux ») mesurant l'écart entre les n_{ij} et les « effectifs théoriques » que l'on aurait si x et y étaient indépendants, c'est-à-dire si les lignes, ou les colonnes du tableau, étaient proportionnelles.

$$\chi^2 = \sum_{i=1}^K \sum_{j=1}^L \frac{(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n})^2}{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}}$$

Valeur théorique
si indépendance

Le coefficient χ^2 est nul dans le cas de l'indépendance (profils identiques), et d'autant plus important que les profils sont différents entre eux.

Vous n'avez pas besoin de savoir ceci !!

$$\chi^2 = \sum_{i=1}^K \sum_{j=1}^L \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}$$

Observée

	Satisfait	Pas satisfait
Homme	70	30
Femme	30	70

théorique si indépendance

	Satisfait	Pas satisfait
Homme	50	50
Femme	50	50

Différences au carré

	Satisfait	Pas Satisfait
Homme	400	400
Femme	400	400

Diviser par le théorique

	Satisfait	Pas satisfait
Homme	8	8
Femme	8	8

Chi² = 32

Représentation graphique entre deux variables qualitatives

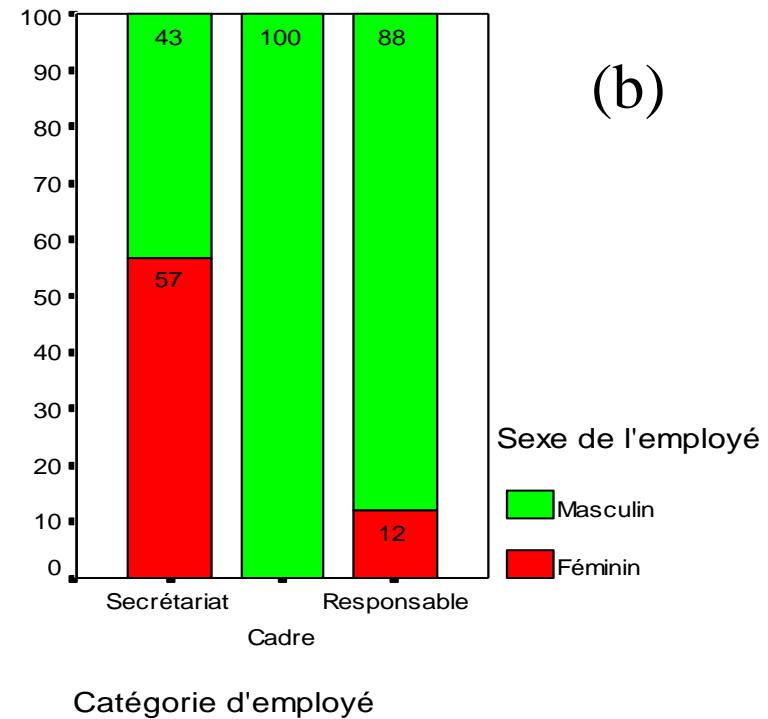
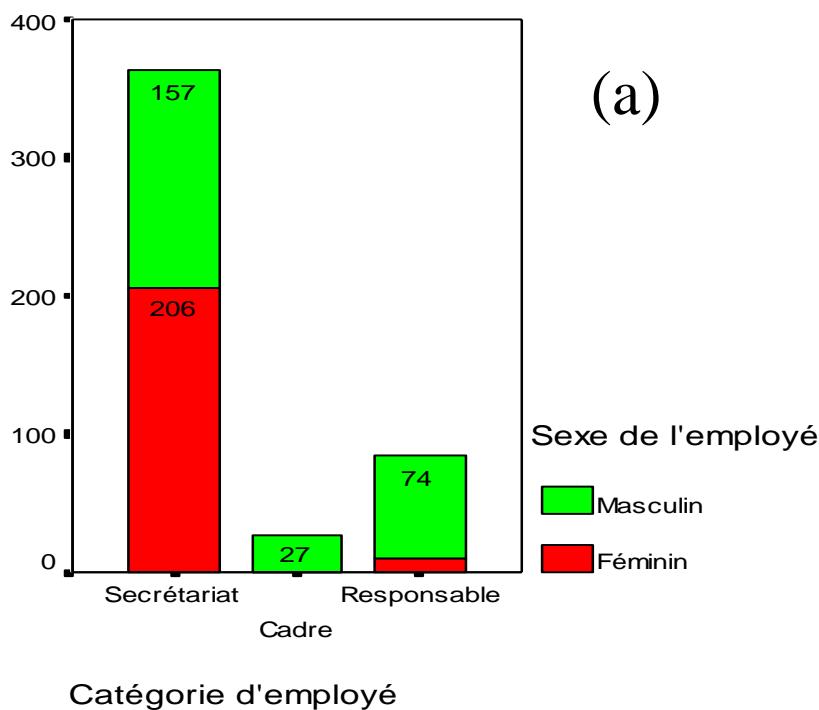


Diagramme en bâtons superposés
(a) : effectifs (b) : pourcentages

Liaisons entre deux variables quantitatives

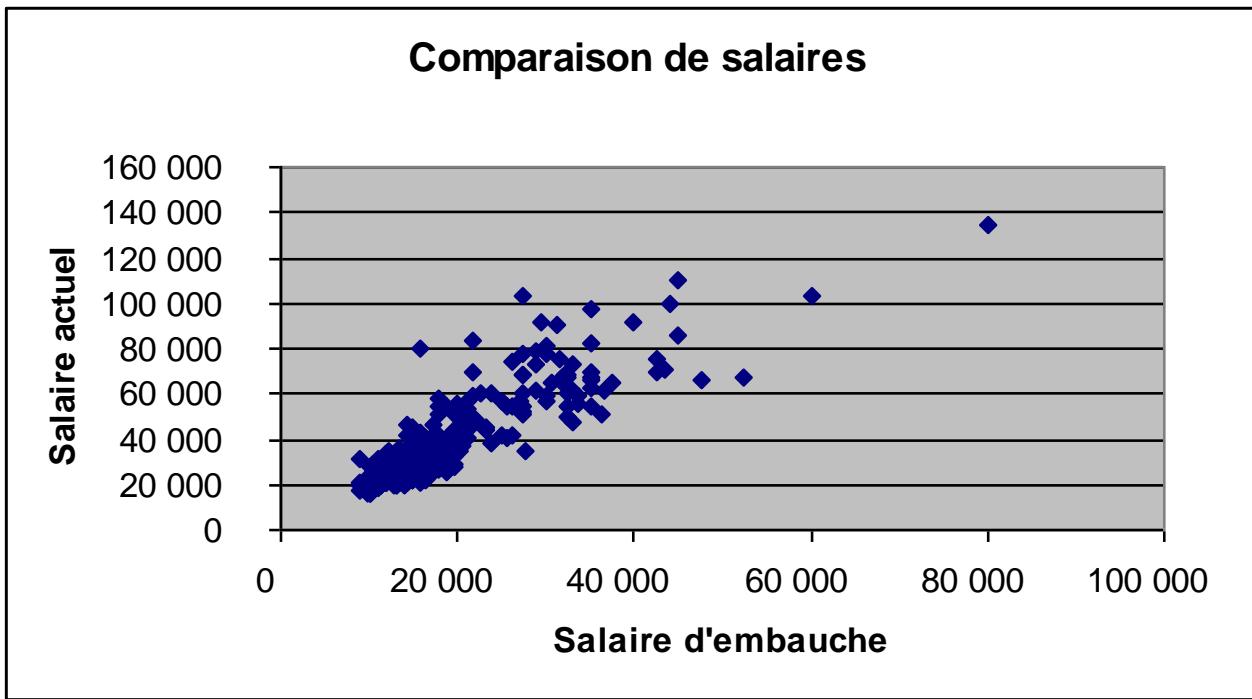
Représentation graphique :

Si l'on dispose de l'observation de deux caractères x et y sur les mêmes n individus, on peut, en plus de l'étude séparée de chaque variable, décrire la liaison entre x et y au moyen d'un **tableau de données brutes** sous forme de n couples de valeurs (x_i, y_i) , $i = 1, \dots, n$.

ID	SALACT	SALDEB
1	\$57 000	\$27 000
2	\$40 200	\$18 750
3	\$21 450	\$12 000
4	\$21 900	\$13 200
5	\$45 000	\$21 000
6	\$32 100	\$13 500
7	\$36 000	\$18 750
8	\$21 900	\$9 750
9	\$27 900	\$12 750
10	\$24 000	\$13 500
11	\$30 300	\$16 500
12	\$28 350	\$12 000
13	\$27 750	\$14 250
14	\$35 100	\$16 800
15	\$27 300	\$13 500
16	\$40 800	\$15 000
17	\$46 000	\$14 250
18	\$103 750	\$27 510
19	\$42 300	\$14 250
20	\$26 250	\$11 550
21	\$38 850	\$15 000
.....		

Liaisons entre deux variables quantitatives

Représentation graphique : Si x et y sont toutes deux variables quantitatives, la représentation graphique consiste en un **nuage de points** M_i de coordonnées (x_i, y_i) , $i = 1, \dots, n$. Sur l'exemple précédent, nous avons :



Liaisons entre deux variables quantitatives

Notion de corrélation : On dit qu'il y a corrélation entre deux variables observées sur des éléments d'une population lorsque les variations des deux variables quantitatives continues se produisent dans le même sens (corrélation positive) ou lorsque les variations sont de sens contraire (corrélation négative).

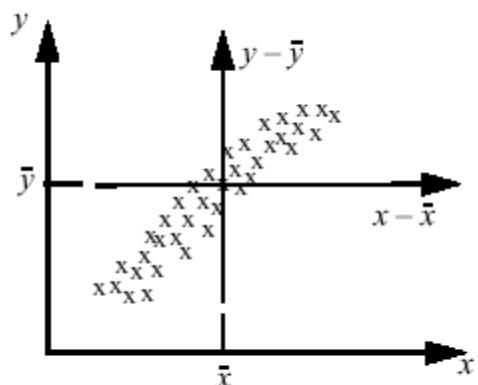
Coefficient de corrélation : noté r est un indice qui rend compte numériquement de la manière dont deux variables quantitatives continues varient simultanément.

$$r_{xy} = \frac{\text{cov}(x,y)}{s_x s_y} \quad \text{où} \quad \text{cov}(x,y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

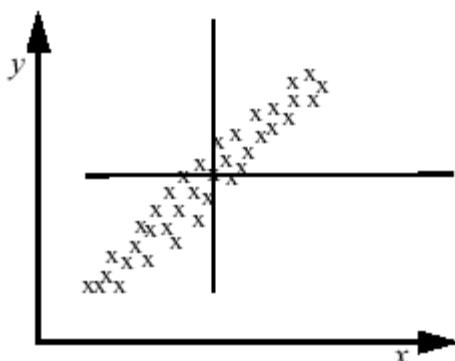
Interpretation du coefficient de corrélation

Coefficient de corrélation :

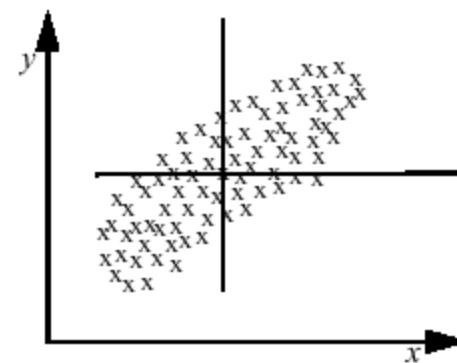
$$r_{xy} = \frac{\text{cov}(x,y)}{s_x s_y}$$



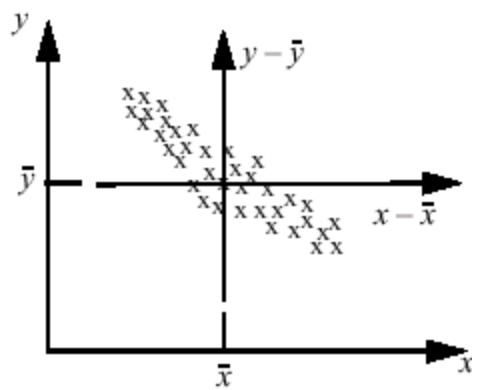
$r > 0$, grand



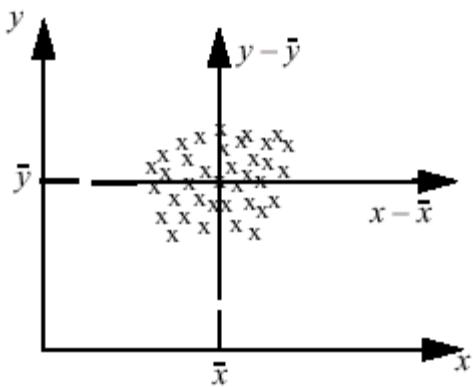
$r \approx 0,9$



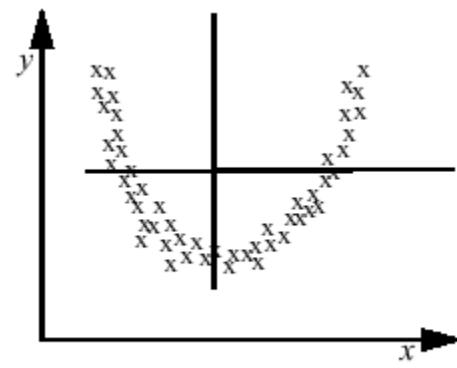
$r \approx 0,5$



$r < 0, |r| \text{ grand}$



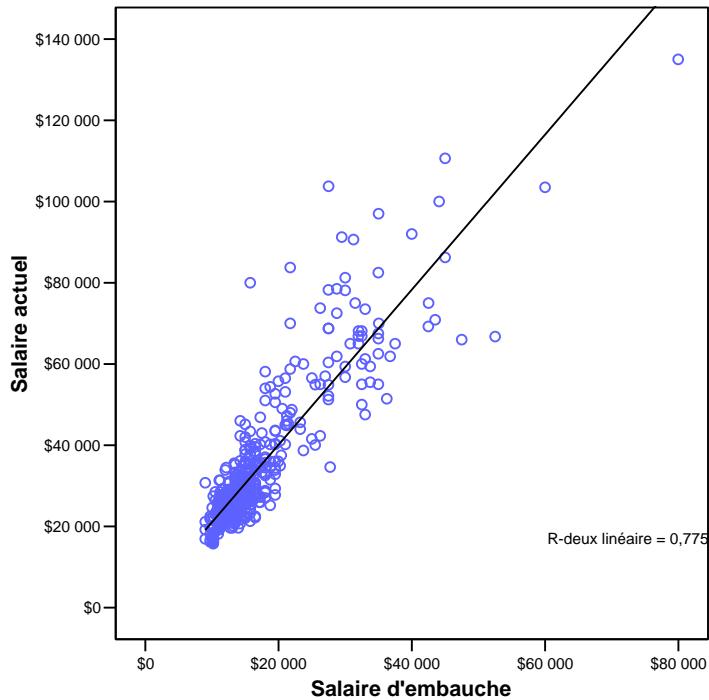
$r \text{ voisin de zéro}$



$r \approx 0$

Exemple de calcul du coefficient de corrélation

Propriétés : r est un coefficient sans unité, indépendant de l'origine choisie, compris entre -1 et $+1$. Il est proche de -1 ou 1 s'il y a une relation presque affine entre x et y .



Matrice de corrélation		
	SALDEB	SALACT
SALDEB	1	
SALACT	0,88	1

Coefficient de corrélation dans l'exemple précédent = 0,88 révèle une forte liaison linéaire entre le salaire actuel et le salaire de début

Liaisons entre une variable qualitative et une variable quantitative

Si x est qualitative à k modalités, l'ensemble des n individus peut être subdivisé en k groupes sur lesquels x est constante.

Si de plus y est quantitative, les k groupes peuvent être représentés par une boîte à moustache de façon à pouvoir les comparer.

Exemples de Liaisons entre variable quantitative et variables qualitatives

■ Analyse de la répartition des salaires

- par sexe
- par niveau d'étude
- par catégories socioprofessionnelle



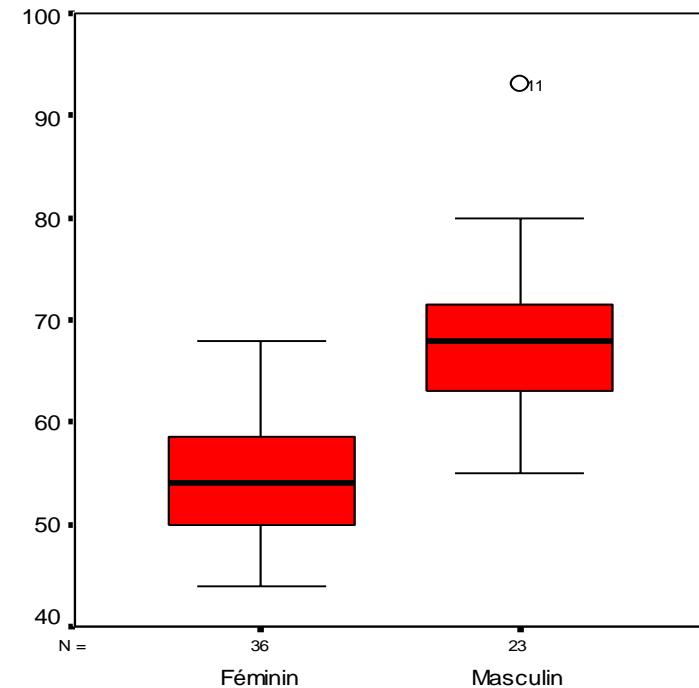
- Différents tableaux croisés 2 à 2 peuvent être générés
- Ou tableaux dynamiques multiples que l'on pourra pivoter selon les différentes dimensions.

Représentation graphique quantitative contre qualitative

Exemple : Comparaison des poids des élèves en fonction du sexe (Féminin, Masculin).

		POIDS				
		Moyenne	Médiane	Q1	Q3	Ecart type
SEXЕ	Féminin	55	54	50	59	6
	Masculin	68	68	62	72	8

Les mesures de position du poids des garçons sont supérieures à ceux des filles. De manière générale, la distribution des garçons est décalées vers le haut par rapport à celle des filles pour une dispersion quasi similaire.



Travail en groupes



Un peu de pratiques :
Séance 1