Cours de statistique descriptive

TD 2 : Les valeurs centrales

Nancy-Université
Université Nancy 2

Lætitia Perrier Bruslé

Introduction et définition des valeurs centrales

- Les valeurs centrales permettent de **résumer** en une seule valeur l'ensemble des valeurs d'une distribution statistique.

- Il existe trois valeurs centrales : le mode, la médiane, la moyenne.

- Les indicateurs de valeurs centrales ne concernent que **les caractères quantitatifs**. Ils s’expriment toujours dans la même unité que celle du caractère

I – La moyenne

Nancy-Université
Université Nancy 2

Lætitia Perrier Bruslé
Cours de statistique descriptive

Introduction

- En théorie, on ne peut calculer la moyenne que pour les caractères quantitatifs continus. Dans la pratique, on la calcule aussi pour des caractères quantitatifs discrets.

- Il existe deux types de moyenne
  - **La moyenne simple** : calculée à partir d'un tableau élémentaire où à chaque élément ne correspond qu'une seule donnée.
  - **La moyenne pondérée** : calculée à partir d'un tableau de dénombrement. Dans ce cas là l'effectif n^i pour la valeur x d'un caractère sert de coefficient pondérateur.

1-1 La moyenne arithmétique simple

- La moyenne est la somme des valeurs divisée par le nombre d'éléments.

$$\overline{X} = \sum_{i=1}^{n} X_i \Big/ N$$

  - N = nombre d'éléments de l'ensemble.
  - Xi = la valeur du caractère X pour un élément i pris au hasard.
  - Σ Xi = Somme des valeurs du caractère X pour la totalité des éléments de l'ensemble.

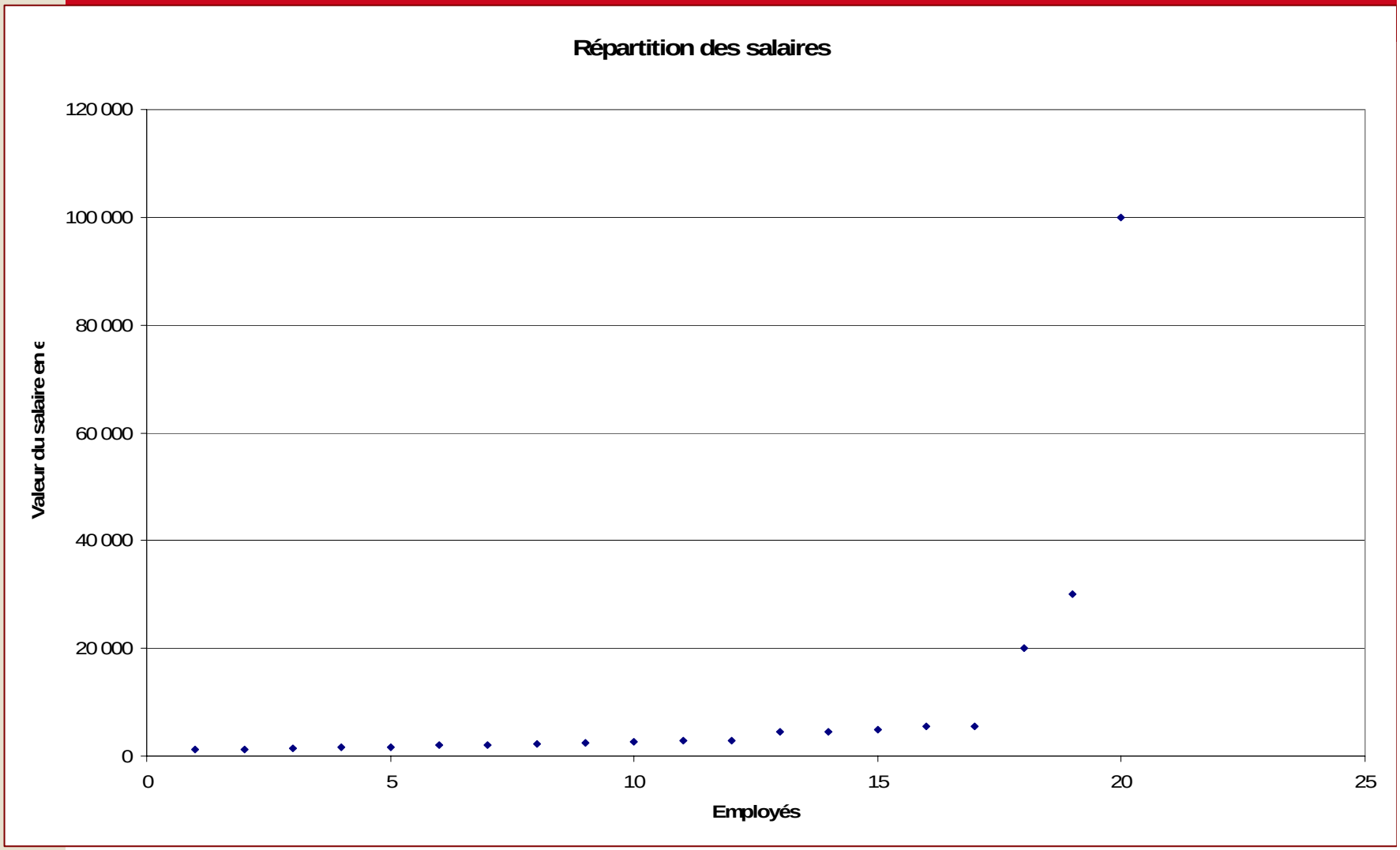
Exemple de moyenne arithmétique simple : Calcul du salaire moyen dans une entreprise

n°	Catégorie	salaire
1	apprentis	1 300
2	apprentis	1 300
3	apprentis	1 500
4	apprentis	1 700
5	apprentis	1 700
6	ouvriers	2 100
7	ouvriers	2 100
8	ouvriers	2 300
9	ouvriers	2 500
10	ouvriers	2 700
11	ouvriers	2 900
12	ouvriers	2 900
13	cadres	4 500
14	cadres	4 500
15	cadres	5 000
16	cadres	5 500
17	cadres	5 500
18	directeur adjoint	20 000
19	directeur adjoint	30 000
20	directeur général	100 000

Question 1 et 2 : Calcul du salaire moyen à partir de différents ensembles

- Salaire moyen 10 000 euros
  - 200 000/20 = 10 000
- Salaire moyen sans le directeur :
  - 100 000/19 = 5263
- Salaire moyen sans le directeur et les sous directeurs
  - 50 000/17 = 2941

La moyenne arithmétique résume très mal la distribution



Valeurs très dispersées (directeur) et forte dissymétrie (concentration des salaires dans les faibles valeurs et dispersion dans les fortes valeurs)

1-2 La moyenne arithmétique pondérée

- Calcul à partir d'un tableau de dénombrement :
  - Regrouper les valeurs du caractère X en classe.
  - Calculer le centre de chaque classe : X_j
  - Pondérer par l'effectif (nombre d'élément de chaque classe) : n_j
  - Diviser la somme des produits $X_j * n_j$ par l'effectif total de l'ensemble statistique : N

$$\overline{X} = \sum_{j=1}^{k} (X_j * n_j) \Big/ N$$

Question 3 : réaliser un tableau de dénombrement à partir du tableau élémentaire

Salaire	Effectif
1 300	2
1 500	1
1 700	2
2 100	2
2 300	1
2 500	1
2 700	1
2 900	2
4 500	2
5 000	1
5 500	2
20 000	1
30 000	1
100 000	1

Tableau de dénombrement
Salaire dans l'entreprise X
en 2006

Définition relative aux classes

- Les classes :
  - Elles correspondent à une partition de l'ensemble de l'intervalle de variation du caractère. (intervalle allant de la valeur minimum prise par le caractère X dans l'ensemble étudié, à la valeur minimum prise par X dans l'ensemble étudié). Elles sont donc définies par une borne supérieure et une borne inférieure.
- Amplitude de la classe
  - Amplitude = Valeur de la borne supérieure – valeur de la borne inférieure
- Centre de la classe *(deux méthodes de calcul)*
  - Centre = Somme de la borne supérieure et de la borne inf. divisée par 2.
  - Centre = Amplitude divisée par deux + borne inf.
- Calcul de la moyenne pondérée
  - On considère que le centre de la classe correspond à la moyenne des individus rassemblés dans cette classe.

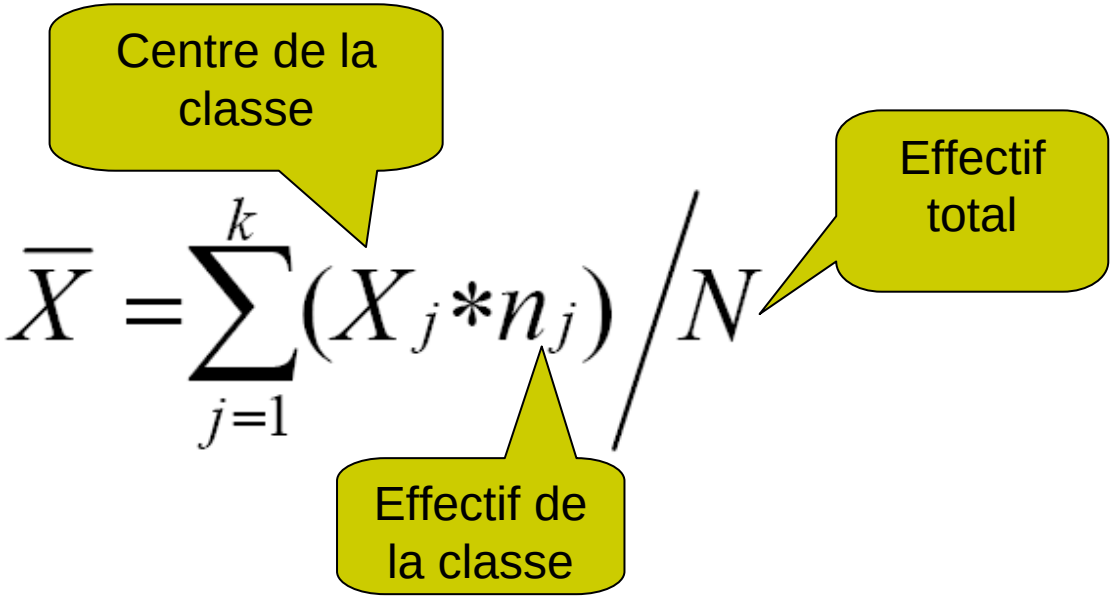
Question 4 : Répartition en classe et calcul de l'amplitude et du centre de chaque classe

Salaires en euro	effectif	Amplitude de la classe	Centre de la classe	Moyenne de la classe
[1000 ; 2000[	5	1000	1 500	1500
[2000, 4000[	7	2000	3 000	2500
[4000; 100 000]	8	96 000	52 000	21875

*Nb : Noter que les classes sont **disjointes** (l'intersection de deux classes est nulle, un élément ne peut appartenir qu'à une seule classe) et **continues** (la partition doit être exhaustive, elle doit intégrer toutes les valeurs que pourrait prendre le caractère dans l'intervalle de variation considéré).*

Question 5 : A partir de ce nouveau tableau de dénombrement calculer la moyenne pondérée.

- Rappel :



- Moyenne pondérée = (5*1500)+(7*3000)+(52000*8)/20 = 22 500

Nb : Cette valeur est beaucoup plus élevée que le salaire moyen réel car les centres des deux dernières classes ne sont pas représentatifs et beaucoup plus élevés que les moyennes des classes auxquelles ils correspondent

Calcul des moyennes pondérées sans tableau de dénombrement

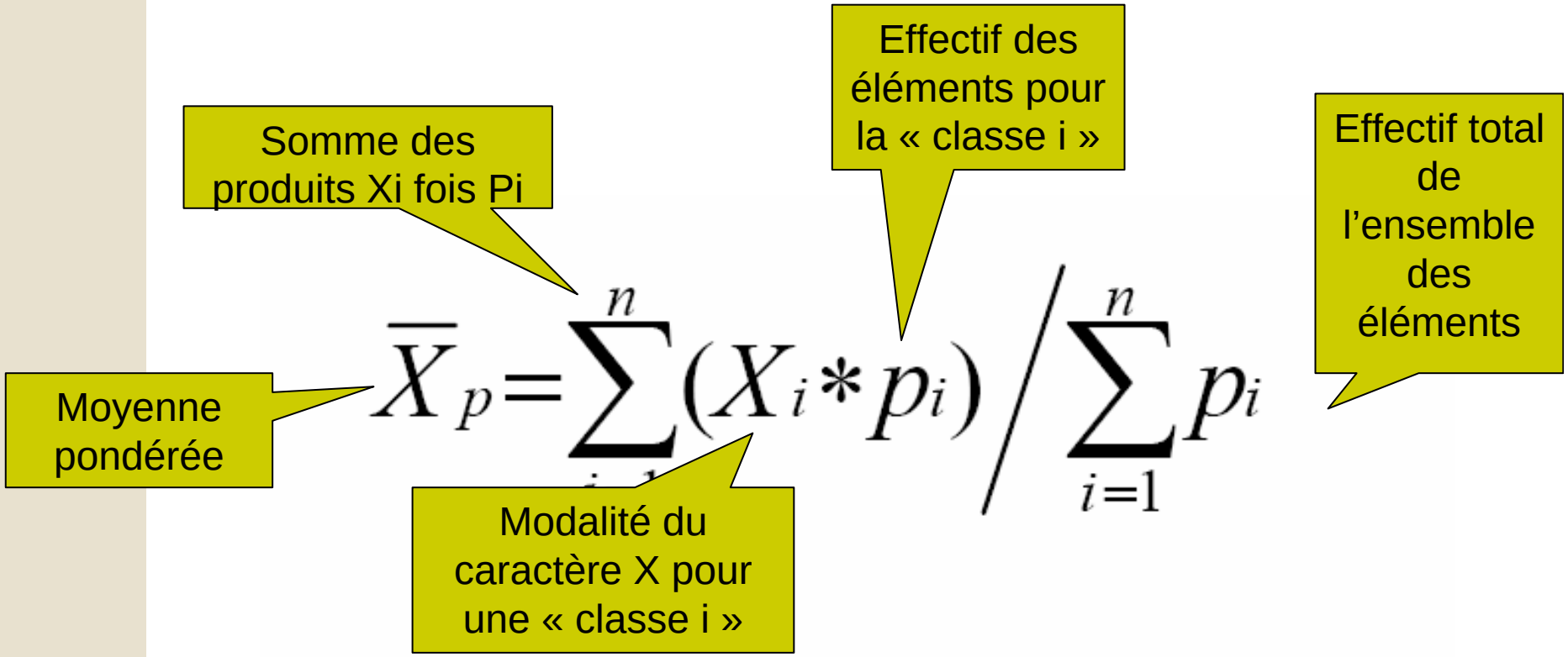
- Souvent les lignes contenues dans un tableau élémentaire correspondent à des ensembles d'individus et non pas à des individus.

- Les modalités du caractère X correspondent alors déjà à des moyennes et la moyenne générale devra pondérer chaque valeur du caractère par le nombre d'individu qu'elle représente.

$$\overline{X}_p = \sum_{i=1}^{n} (X_i * p_i) \Big/ \sum_{i=1}^{n} p_i$$

Calcul des moyennes pondérées sans tableau de dénombrement

- Calcul de la moyenne pondérée dans ce cas :



Calcul des moyennes pondérées sans tableau de dénombrement

- Souvent les lignes contenues dans un tableau élémentaire correspondent à des ensembles d'individus et non pas à des individus.

- Les modalités du caractère X correspondent alors déjà à des moyennes et la moyenne générale devra pondérer chaque valeur du caractère par le nombre d'individu qu'elle représente.

$$\overline{X}_p = \sum_{i=1}^{n} (X_i * p_i) \Big/ \sum_{i=1}^{n} p_i$$

Exemple : calculer la moyenne pondérée à partir de ce tableau de dénombrement

Catégorie	salaire moyen en euro (Xi)	Effectif (Pi)
apprentis	1 500	5
ouvriers	2 500	7
cadres	5 000	5
directeurs adjoints	25 000	2
directeur général	100 000	1
Total masse salariale	200 000	20

- Il faut pondérer le salaire moyen de chaque catégorie de salariés par son effectif.

- Moyenne pondérée = (5*1500 + 7*2500 + 5*5000 + 2*25000 +1*100 000) / 20 = 10 000 euros.

1-3 Calcul sur les moyennes : moyenne des taux et taux moyen

- Une application directe de la moyenne pondérée concerne les caractère quantitatif de taux,
  - < caractères X définis comme le rapport de deux caractères de stock V (numérateur) et P (dénominateur).
- Lorsque l'on considère un ensemble de N d'éléments décrits par le caractère X, il faut clairement distinguer le taux moyen et la moyenne des taux.
- Taux moyen = La valeur du rapport V/P si tous les individus étaient fusionnés
- Moyenne des taux = La moyenne des valeurs d'individus de poids différents

Exemple d'application : calculer la moyenne des taux et la moyenne pondérée

Pays	**PIB ($ / hab.)** *- Caractère de taux -*	**Population en million d'habitants**
Chine Populaire	3 500	1300
Taïwan	18 500	20

Moyenne des taux	(18 500 + 3500)/2 = 11 000

Taux moyen = Moyenne pondérée	(3500*1300) + (18500*20)/1320 = 3727

Question 6 : description du tableau

- Ensemble : les régions françaises.
- Elément : chacune des régions.
- Population en milliers : quantitatif, mesurable, de stock, discret
- Bac +2 : quantitatif, mesurable, de taux, continu
- Zone : qualitatif, nominal
- Moyenne des taux = moyenne régionale métropolitaine
  - Additionner les valeurs de X pour le caractère « Bac+2 » et divisé par l'effectif (=nombre de région = 22) : 343/22=15,63

Question 7 Moyenne des régions et moyenne française

- Moyenne des régions (moyenne des taux) :
  - 15,6 (moyenne des modalités du caractère Bac+2 pour les 22 régions).
- Moyenne française (taux moyen)
  - 18,4 (hommes ayant un niveau bac+2 rapportée à la population masculine)
  - Pour la calculer : faire la somme des produits (taux Bac+2*population régionale)/population totale française
  - De cette façon : chaque taux est pondéré par le poids de la population régionale

Question 8 : Moyenne par sous ensemble régional

zone NNE	Pop° 2003	Bac+2 ans
Champagne Ardennes	1 337	12,6
Bourgogne	1 612	13,1
Picardie	1 869	13,2
Hte Normandie	1 787	13,9
Centre	2 467	14
Lorraine	2 319	14,5
Franche Comté	1 131	14,7
Nord-Pas de Calais	4 013	14,9
Alsace	1 775	18,1
PACA	4 665	18,7
Rhône Alpes	5 814	19,2
Île-de-France	11 131	30,1
Moyenne des taux		**16,4**
Moyenne sans IDF		**15,2**
Taux moyen		19,9

ZONE SSO	Pop° 2003	Bac+2ans
Limousin	711	12,6
Bse Normandie	1 436	12,7
Poitou-Charente	1 668	12,7
Auvergne	1 314	13,4
Corse	266	13,4
Pays-de-Loire	3 312	14,7
Bretagne	2 978	16
Aquitaine	2 988	16,3
Languedoc Roussillon	2 402	16,8
Midi-Pyrénées	2 638	18,2
Moyenne des taux		**14,68**

Conclusion : propriété de la moyenne

- La somme des écarts à la moyenne est égale à zéro

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)=0$$

- La moyenne minimise les distances au carré

$$\sum_{i=1}^{n}\left(X_i-A\right)^2$$

est minimum si ,et seulement si, A est la moyenne du caractère X

II – La médiane

Nancy-Université
Université Nancy 2

Lætitia Perrier Bruslé
Cours de statistique descriptive

Introduction

- La médiane ne peut-être calculée que pour **les caractères quantitatifs**. Les éléments étant classés en fonction des valeurs du caractère par ordre croissant.

- La médiane est la valeur du caractère qui partage l'ensemble statistique **en deux ensembles d'effectifs égaux** : 50 % des valeurs lui sont supérieures et 50 % lui sont inférieures.

- La médiane n'est qu'une forme particulière de **fractile** (appelés aussi quantile).

  - Les fractiles sont **des paramètres de position**. Ils divisent la distribution en un certain nombre de parties égales en fonction du nombre d'individus et non pas en fonction de leur valeur

Calcul de la médiane à partir du tableau élémentaire

- Ordonner le tableau et repérer l'élément qui partage la distribution en deux parties égales, c'est à dire celui qui a le rang (n+1)/2.

- Deux cas de figure sont possibles : Soit la distribution a un nombre impair d'élément on trouve une valeur unique qui est la médiane.

  - Ex : si 29 éléments = (29+1)/2 = 15 : la médiane correspond à la valeur du caractère X pour l'élément au 15ème range

- Soit la distribution a un nombre pair d'élément, on trouve deux valeurs qui déterminent un **intervalle médian** : on prend alors pour médiane le centre de cet intervalle médian.

  - Ex : si la distribution compte 28 éléments : (28+1)/2 = 14,5. La médiane correspond à l'intervalle médian entre la valeur de l'élément au rang 14 et la valeur de l'élément au rang 15.

Application : dans l'entreprise X, quel est le salaire médian ?

Rang	Catégorie	salaire
1	apprentis	1 300
2	apprentis	1 300
3	apprentis	1 500
4	apprentis	1 700
5	apprentis	1 700
6	ouvriers	2 100
7	ouvriers	2 100
8	ouvriers	2 300
9	ouvriers	2 500
10	ouvriers	2 700
11	ouvriers	2 900
12	ouvriers	2 900
13	cadres	4 500
14	cadres	4 500
15	cadres	5 000
16	cadres	5 500
17	cadres	5 500
18	directeur adjoint	20 000
19	directeur adjoint	30 000
20	directeur général	100 000

- Distribution avec un nombre pair d'élément
- Rang de la médiane =
  - (20+1)/2 = 10,5
- Intervalle médian
  - Entre 2700 et 2900
- Médiane : centre de l'intervalle
  - (2900+2700)/2 = 2800

Rang	Nom	Bac+2ans
1	**Champagne Ardennes**	12,6
2	**Limousin**	12,6
3	**Bse Normandie**	12,7
4	**Poitou-Charente**	12,7
5	**Bourgogne**	13,1
6	**Picardie**	13,2
7	**Auvergne**	13,4
8	**Corse**	13,4
9	**Hte Normandie**	13,9
10	**Centre**	14
11	**Lorraine**	14,5
12	**Franche Comté**	14,7
13	**Pays-de-Loire**	14,7
14	**Nord-Pas de Calais**	14,9
15	**Bretagne**	16
16	**Aquitaine**	16,3
17	**Languedoc Roussillon**	16,8
18	**Alsace**	18,1
19	**Midi-Pyrénées**	18,2
20	**PACA**	18,7
21	**Rhône Alpes**	19,2
22	**Île-de-France**	30,1

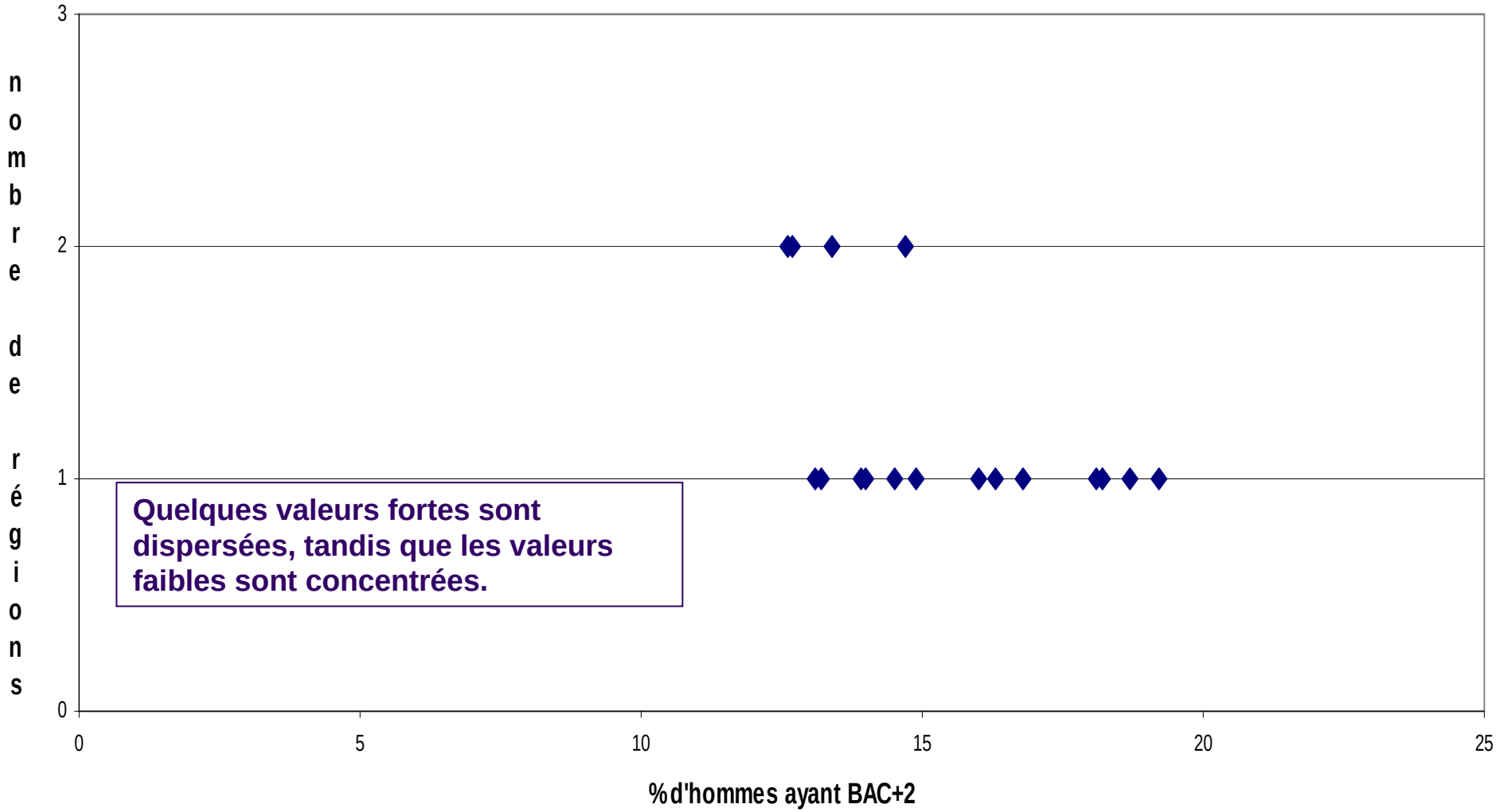
Question 10 : Tableau ordonné en fonction du caractère

- Distribution avec un nombre pair d'élément
- Rang de la médiane =
  - (22+1)/2 = 11,5
- Intervalle médian
  - Entre 14,5 et 14,7
- Médiane : centre de l'intervalle
  - (14,5+14,7)/2 = 14,6

Question 10 Réflexions sur la valeur de la médiane

- Comparaison moyenne – médiane
  - Médiane = 14,6
  - Moyenne = 15.6
  - La médiane est inférieure à la moyenne.
- Explication : La moyenne est plus élevée car elle tient compte des valeurs exceptionnelles.
- Le diagramme de distribution en témoigne.

Diagramme de distribution du niveau de qualification des hommes de plus de 25 ans dans la population française



Règle générale

- Lorsqu'il y a une dissymétrie marquée de la distribution statistique, la médiane est généralement préférable à la moyenne car elle est moins influencée par les valeurs exceptionnelles qui sont à l'origine de la dissymétrie.

- C’est seulement lorsque la dissymétrie est peu marquée, qu’on pourra utiliser la moyenne.

Exercice complémentaire

- Réorganiser l'ensemble en deux tableaux : population du SSO et population du NNE.
- Calculer la médiane et la moyenne pour ces deux ensembles statistiques.
- En tirez une conclusion.

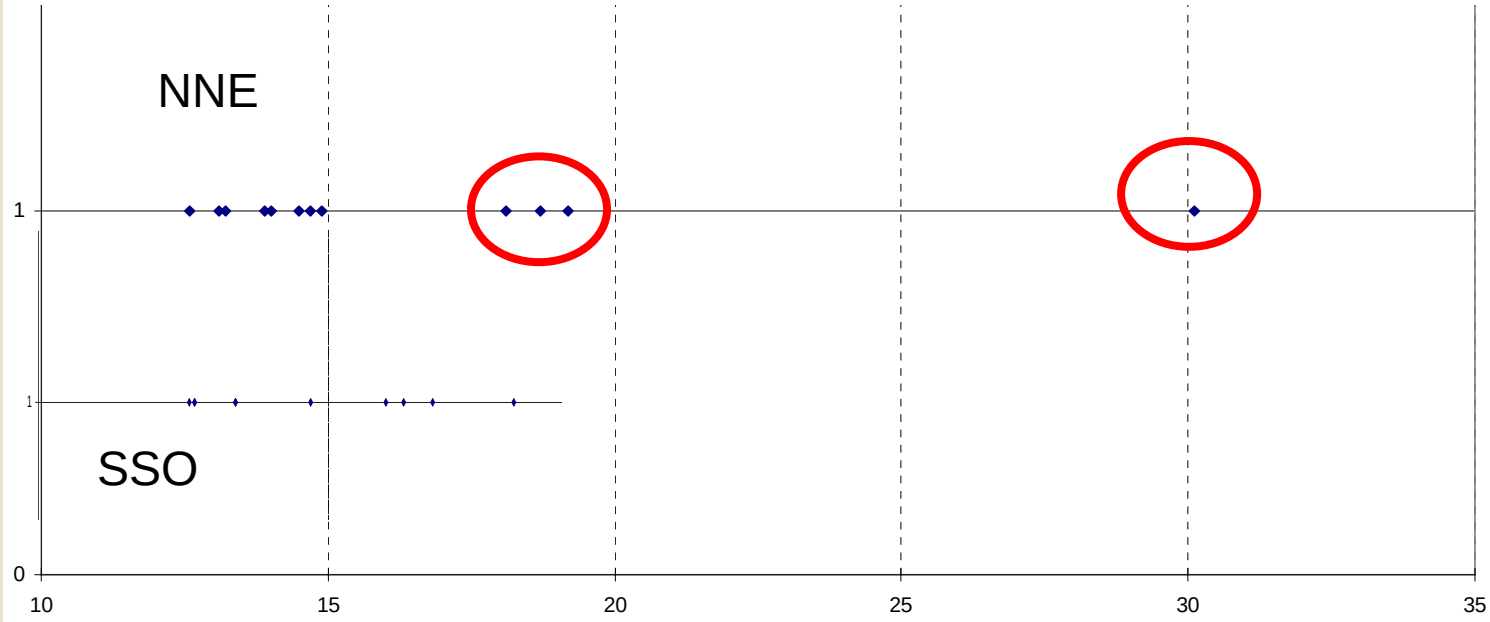
Exercice complémentaire

- Réorganiser l'ensemble en deux tableaux : population du SSO et population du NNE.
- Calculer la médiane et la moyenne pour ces deux ensembles statistiques.
- En tirez une conclusion.

	ZONE SSO	Bac+2ans
1	Limousin	12,6
2	Basse Normandie	12,7
3	Poitou-Charentes	12,7
4	Auvergne	13,4
5	Corse	13,4
6	Pays de Loire	14,7
7	Bretagne	16
8	Aquitaine	16,3
9	Languedoc Roussillon	16,8
10	Midi-Pyrénées	18,2
11		
	Moyenne	14,68
	Médiane	**14,05**

	Zone NNE	Bac+2ans
1	Champagne Ardennes	12,6
2	Bourgogne	13,1
3	Picardie	13,2
4	Hte Normandie	13,9
5	Centre	14
6	Lorraine	14,5
7	Franche Comté	14,7
8	Nord Pas de Calais	14,9
9	Alsace	18,1
10	PACA	18,7
11	Rhône Alpes	19,2
12	Île-de-France	30,1
	Moyenne	**16,41**
	Médiane	**14,6**

Comparaison des deux distributions



III – Le mode

Nancy-Université
Université Nancy 2

Lætitia Perrier Bruslé
Cours de statistique descriptive

Le mode : notions de base

- Le mode (parfois appelé valeur dominante) est la valeur de la variable statistique quantitative pour laquelle l'effectif est le plus élevé.
- Il se calcule toujours à partir d'un dénombrement de l'effectif pour chaque modalité du caractère.
- Il faut donc partir d'un tableau de dénombrement pour le déterminer.

Calcul du mode pour les variables discrètes

- Pour un caractère qualitatif ou quantitatif discret ayant un nombre de modalités inférieur au nombre d'éléments, le mode est la modalité qui a la fréquence simple la plus élevée (ou l'effectif le plus élevé, ce qui revient au même).

- La fréquence simple correspond à la proportion d'individus associés à une modalité particulière d'un caractère statistique. Elle permet de cerner le poids respectifs des modalités d'un caractère.

Application

- Exemple : Dans l'entreprise X quelle est la catégorie de salariés la plus représentée ?

- Réponse : il y a 5 apprentis, 7 ouvriers, 5 cadres, 2 sous-directeurs et 1 directeur. La catégorie modale est donc celle des ouvriers.

Catégorie	Effectif
apprentis	5
ouvriers	7
cadres	5
directeurs adjoints	2
directeur général	1

Calcul du mode pour les variables continues

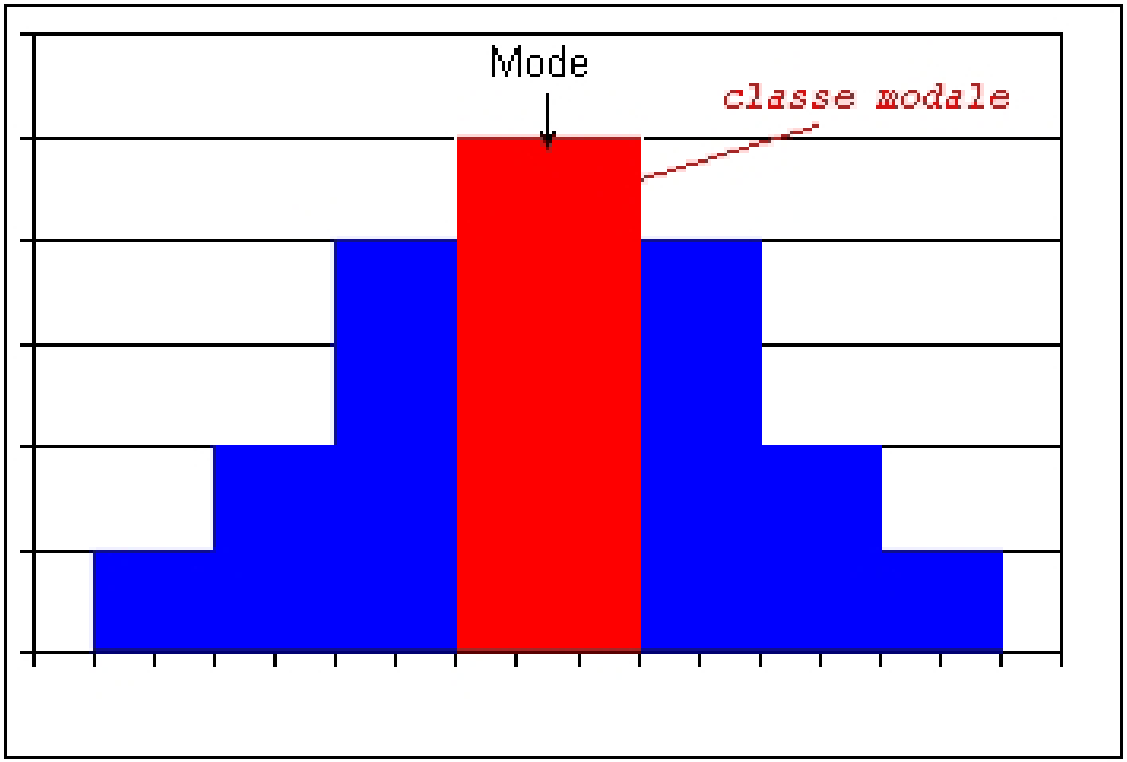
- Les modalités sont en nombre infini. Il est peu probable que deux éléments aient la même valeur.

- Il faut donc partir des classes pour définir le mode.

  ➲ Faire une partition en classes de la distribution statistique.

- Le mode est alors **le centre de la classe modale**, c'est à dire la classe qui a la **fréquence moyenne** la plus élevée.

Le mode est le centre de la classe modale, c'est à dire la classe qui a la fréquence moyenne la plus élevée



Fréquence simple

- ***Fréquence simple Fi*** d'une modalité ou d'une valeur Xi est le rapport entre l'effectif correspondant à cette modalité et l'effectif total de la distribution. La fréquence varie de 0 à 1 elle est alors notée sous forme décimale dans [0;1]. Elle peut être exprimée en pourcentage, elle varie alors de 0% à 100%.
  - fi = ni / N (fréquence sous forme décimale)
  - fi =100 * fi / N (fréquence en pourcentage)
- La somme des fréquences simples est égale à 1 (ou à 100 %) des éléments.

Fréquence moyenne

- ***La fréquence moyenne*** ou densité d'effectif mesure la concentration des éléments à l'intérieur d'une classe.

- Cette valeur est toujours calculée à une constante près de sorte qu'il est indifférent d'utiliser l'une ou l'autre des formules suivantes.

- *Fmj = Fj / Aj* ou *Fmj = Ej / Aj*

  - Fmj = Fréquence moyenne de la classe j
  - Fj = Fréquence simple
  - Aj = Amplitude de classe
  - Ej = Effectif de la classe

Pour l'entreprise Z calculer la fréquence simple et la fréquence moyenne de chaque classe

Salaires	Effectif de la classe Ni
[1000 ; 2000[	5
[2000, 4000[	7
[4000; 100000]	8

Fréquence simple *Ni/N*	Amplitude de la classe *=BorneSup-borne inf.*	Fréquence moyenne de la classe = Fréquence simple / Amplitude
25%	1000	0.00500
35%	2000	0.00350
40%	96000	0.00008

Attention

- La fréquence moyenne est différente de la fréquence simple.

- Donc la classe modale n'est pas forcément celle qui a l'effectif (c'est à dire la fréquence simple) la plus élevée : c'est la classe où les éléments sont le plus concentrés (c'est à dire celle où la fréquence moyenne est la plus élevée)

Question 11 : calcul du mode de la distribution pour le tableau 2

- Le caractère : hommes ayant au moins le niveau bac plus 2 : est un caractère continu.

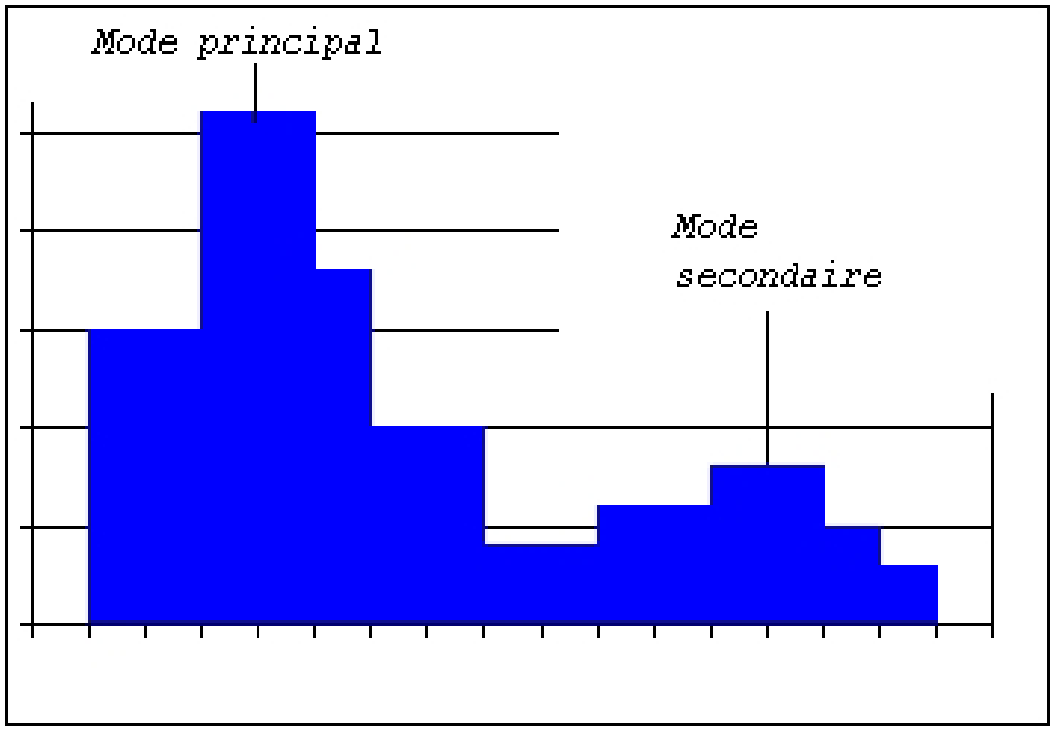
- Il faut donc construire un tableau de dénombrement regroupé en classe pour établir le mode.

Classe	Effectif	Fréquence simple	Amplitude	Fréquence moyenne
[12,6;13,2]	6	0,27	0,6	10
]13,2;14]	4	0,18	0,8	5
]14;18]	4	0,18	4	1
]18;30,1]	8	0,36	12,1	0,66

Utilisation du mode pour caractériser les distributions statistiques

- **Distribution unimodale :** Un seul mode.

- **Distributions bimodales ou multimodales :** La distribution comporte plusieurs modes.



Mode et type de distribution

- Distribution unimodale symétrique
  - Lorsque la distribution est unimodale et symétrique, on va trouver à peu près moyenne = médiane = mode.
  - Le meilleur résumé est alors donné par la moyenne car elle tient compte de toutes les observations et elle possède des propriétés statistiques intéressantes.
- Distributions unimodales dissymétriques :
  - Concentration pour les valeurs faibles, ou concentration pour les valeurs fortes.
- Distribution bi-modale
  - Le mode principal est différent de la moyenne et de la médiane qui ont de fortes chances de correspondre à une zone de dispersion des valeurs. Dans ce cas, ni la moyenne ni la médiane ne sont significatives.

Conclusion : intérêt du mode

- La détermination du mode pour un caractère continu rend obligatoire l'établissement d'une partition en classe. Pour une même distribution, on peut avoir des modes différents selon le découpage en classe qui a été choisi.

- Le mode est donc une valeur centrale qui est assez fragile pour les caractères continus. En revanche, c'est la seule valeur centrale possible pour les caractères qualitatifs.