Statistique : le terme statistique désigne à la fois :

1) l'ensemble des données numériques concernant une catégorie de faits (sens très ancien). Il s'agit de l'expression dans sa signification la plus usuelle (ex. "la statistique du chômage en 1995")
2) l'ensemble des méthodes mathématiques permettant :
   a) de résumer quantitativement l'information recueillie sur un ensemble d'éléments au moyen d'une investigation exhaustive. C'est la **statistique descriptive**, qui fait l'objet de ce cours.
   b) de généraliser à de grands ensembles d'éléments les conclusions tirées des résultats obtenus avec des ensembles beaucoup plus restreints appelés échantillons. C'est la **statistique inférentielle** ou probabiliste, qui n'est pas abordée dans cette UE

1. Vocabulaire : Introduction au tableau élémentaire

Ensemble : c'est la collection (le plus souvent finie en géographie) d'unités, ou d'éléments, sur laquelle porte l'observation. Pour que cet ensemble soit correctement défini, il faut lui donner une définition précise de façon à ce que deux personnes différentes aboutissent toujours à la même liste d'éléments. L'ensemble des éléments observés sera appelé E.

Elément : les éléments sont les objets constitutifs de l'ensemble. Ce sont des objets déterminés dont l'appartenance à tel ou tel ensemble E est sans ambiguïté. Les éléments peuvent être désignés par leur position dans le tableau de données : 1 pour le premier, i pour un élément quelconque, N pour le dernier élément.

Caractère : les éléments d'un ensemble sont décrits par un caractère. Cela revient à établir une correspondance entre chaque élément i de l'ensemble E et l'ensemble X des modalités ou des valeurs du caractère. La fonction f E --> X i --> x_i est une application au sens mathématique: chaque élément de E a une modalité (caractère qualitatif) ou une valeur (caractère quantitatif) et une seule dans X.

Modalité, Mesure : les différentes situations où les éléments de E *peuvent se trouver* à l'égard d'un caractère qualitatif considéré, sont les différentes **modalités** du caractère qualitatif X. Dans le cas ou le caractère X est quantitatif, les différentes situations où les éléments de E peuvent se trouver sont des **mesures**. Ces *modalités* ou ces *mesures* doivent être à la fois incompatibles (un élément de E ne peut prendre qu'une seule modalité) et exhaustive (à chaque élément de E doit pouvoir correspondre une modalité de X) de sorte que chaque élément de E ait une modalité et une seule dans X.

Tableau élémentaire : c'est un tableau à simple entrée où les lignes correspondent aux éléments de l'ensemble étudié et les colonnes aux caractères (ou variables) décrivant ces éléments (figure 1). La première colonne est en principe réservée à la liste nominale des éléments.

Types de caractères

caractère qualitatif : les *modalités* ne sont pas mesurables, ce sont des noms ou ce qui revient au même des sigles ou des codes. Les différentes modalités ne sont pas ordonnables. Attention, même si les modalités sont des codes numériques, les opérations sur les modalités n'ont aucun sens. exemple: type de relief avec trois modalités (plaine, montagne, plateau), ou encore taille de la ville avec quatre modalités (petite, moyenne, grande, très grande

Caractère ordinal: il est exprimé sur une échelle ordinale: chaque *modalité* est explicitement significative du rang pris par chaque individu pour le caractère considéré. Si E possède N éléments, les modalités seront 1^{er}, 2^{eme}, 3^{eme}, ...n^{eme}. Comme on possède juste l'ordre des individus, on ne sait rien de l'intervalle des valeurs. Exemple: Rang des départements pour la population en 1999.

Caractère quantitatif : les différentes situations où peuvent se trouver les éléments sont des *mesures*; elles sont ordonnables et la moyenne a une signification

*** repérable ou mesurable**

- *quantitatif repérable sur une échelle d'intervalle*. Ces caractères permettent de repérer la position de chaque élément par rapport à une origine arbitraire. La valeur 0 est donc conventionnelle et ne signifie pas l'absence du phénomène. Exemple: Latitude, longitude, température, altitude, ...
- *quantitatif mesurable sur une échelle numérique* . Le 0 signifie bien l'absence du phénomène Exemple: population, taux de fécondité, précipitations

***stock ou taux**

- *les caractères quantitatifs de stock* expriment des quantités concrètes : la somme des valeurs prises par l'ensemble E des éléments a un sens. Exemple: Population totale d'un département.

- *les caractères quantitatifs de taux* expriment le rapport entre deux valeurs, on les appelle parfois caractères de rapport. Leur total n'a pas de signification. Exemple: Densité de population d'un département , ou proportion des actifs chômeurs à une date donnée.

*** discret ou continu**

- *les caractères quantitatifs discrets* sont des caractères dont les différentes situations où peuvent se trouver les éléments sont des nombres isolés dont la liste peut être établie a priori. Exemple: Nombre de villes de plus de 100 000 h dans chaque département.

- *les caractères quantitatifs continus* sont des caractères les différentes situations où peuvent se trouver les éléments ne sont pas dénombrables et qui sont définies sur un intervalle (continu) de valeur donnée. Exemples: Superficie des départements, Altitude moyenne des départements, salaire moyen des employés en 1999

2. Vocabulaire : Résumer une distribution statistique par une mesure : les valeurs centrales

Le but des valeurs centrales est de résumer en une seule valeur l'ensemble des valeur d'une distribution statistique. Il existe trois valeurs centrales : le mode, la médiane, la moyenne.

Le mode

Le mode, ou **valeur dominante,** est la valeur la plus fréquente d'une distribution. Cette valeur se calcule toujours à partir d'un dénombrement des modalités du caractère. Il faut donc distinguer le cas des caractères discrets et des caractères continus (Cf. Vocabulaire 2).

* caractère qualitatif et caractère discret

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

* Caractère quantitatif continu

Les modalités étant en nombre infini, il est peu probable que deux éléments aient la même valeur. Dans ce cas, le mode ne peut pas être défini directement, il faut au préalable établir une partition en classes. Le mode est alors le centre de la classe modale, c'est à dire de la classe qui a **la fréquence moyenne la plus élevée**.
Le mode correspond à la valeur lue en abscisse du sommet de l'histogramme. Lorsque celui-ci présente deux pics séparés par un creux, on dit que la distribution est **bimodale**.

La médiane

Les *valeurs du caractère X étant classées par ordre croissant*, la médiane est la valeur du caractère qui partage l'ensemble décrit par X en deux sous ensembles d'effectifs égaux : 50 % des éléments ont des valeurs de X supérieures à $X_{méd}$ et 50% prennent des valeurs inférieures. La médiane ne peut être calculée que pour les caractères quantitatifs.

- Calcul à partir du tableau élémentaire :

On ordonne le tableau, et on cherche l'élément qui partage la distribution en deux parties égales: on repère l'élément qui a le rang (N+1)/2 pour le caractère X. Si la distribution a un nombre impair d'éléments on trouve une valeur unique qui est la médiane, si la distribution a un nombre pair d'éléments, on trouve deux valeurs qui déterminent un **intervalle médian** : on prend alors pour médiane le centre de cet intervalle médian.

- Calcul à partir de la courbe des fréquences cumulées :

(Cf. Vocabulaire 2)

- Calcul à partir d'un tableau de dénombrement :

On repère la classe j qui contient la médiane, puis on réalise une interpolation linéaire pour estimer la valeur de celle-ci selon la formule :

Médiane = B_{infj} + [A_j/ F_j* (0.5-F_{ascj-1})]

- Propriétés de la médiane

La médiane est la valeur du caractère qui est la plus proche de toutes les autres. C'est celle qui minimise les distances en valeur absolue :

$\sum_{i=1}^{N} |x_i - x_{méd}|$ est minimum si et seulement si $x_{méd}$ est la médiane du caractère X

La moyenne

Elle est calculée pour les caractères quantitatifs

- **calcul à partir du tableau élémentaire:**

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

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

- **calcul à partir du tableau de dénombrement:**

On effectue une moyenne pondérée en assimilant chaque classe j à son centre X_j et en pondérant par l'effectif n_j de la classe.

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

- **moyenne pondérée :**

Plus généralement, on recourt à la pondération lorsque les unités n'ont pas le même poids. Si chaque unité i est décrite par sa modalité xi et son poids pi, la moyenne pondérée est :

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

- **Propriétés de la moyenne**

1) Si A = **moyenne de X**

$$\sum_{i=1}^{n} X_i = n * \overline{X}$$

2) La somme des écarts à la moyenne est égale à zéro.

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

–

3) La moyenne minimise les distances au carré

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

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

Avantages et inconvénients des différentes valeurs centrales :

Le statisticien Yule (XIXème siècle) a défini six propriétés souhaitables pour les valeurs centrales. Le tableau ci-dessous permet de montrer les avantages et inconvénients des trois valeurs centrales (propriété + réalisée, - non réalisée

Propriétés	**Mode**	**Médiane**	**Moyenne**
1) est définie de façon objective	+	+	+
2) dépend de toutes les valeurs observées	-	-	+
3) a une signification concrète	+	+	-
4) est simple à calculer	+	+	+
5) est peu sensible aux fluctuations de l'échantillon	-	+	-
6) se prête au calcul algébrique	-	-	+

Vocabulaire 3. Evaluer le degré de dispersion des valeurs d'une distribution statistique

Dispersion statistique : On appelle dispersion statistique, la tendance qu'ont les valeurs de la distribution d'un caractère à s'étaler, à se disperser, de part et d'autre d'une valeur centrale. On distingue la dispersion absolue (mesurée dans l'unité de mesure du caractère) et la dispersion relative (mesurée par un nombre sans dimension).

LES PARAMETRES DE DISPERSION ABSSOLUE

Les paramètres de dispersion absolue indiquent de combien les valeurs d'une distribution s'écartent en général de la valeur centrale de référence. Un paramètre de dispersion absolue s'exprime toujours dans l'unité de mesure de la variable considérée. Les quatre paramètres de dispersion absolue les plus courants sont l'étendue, l'intervalle inter quantile, l'écart absolu moyen et l'écart type.

1) Etendue : l'étendue d'une distribution est égale à la différence entre la plus grande et la plus petite valeur de la distribution :

Etendue de X = X_{max} - X_{min}

2) Mesures de la dispersion statistique en référence à la médiane

Quantiles : on appelle quantiles les valeurs du caractère qui définissent les bornes d'une partition en classes d'effectifs égaux.

- Les **quartiles** sont les trois valeurs qui permettent de découper la distribution en quatre classes d'effectifs égaux. On les note Xq1 , Xq2 et Xq3.

Partition du caractère X_{min} → X_{q1} → X_{q2} → Q_{q3} → X_{max}
fréquence des éléments 25% 25% 25% 25%

Remarque : X_{q2} est égal à la médiane.

- **L'intervalle interquartile** est l'étendue de la distribution sur laquelle se trouvent concentrée la moitié des éléments dont les valeurs de X sont les plus proches de la médiane. On exclut alors de la distribution les 25% des valeurs les plus faibles et les 25 % des valeurs les plus fortes de X. Cet intervalle se note:(X_{q3}-X_{q1}).

- Les **déciles** sont les neufs valeurs de X qui permettent de découper la distribution en dix classes d'effectifs égaux. 0n les note X_{d1}...X_{d9}.

- L'intervalle inter décile est l'étendue de la distribution sur laquelle se trouvent concentrés 80% des éléments dont les valeurs de X sont les moins différentes de la médiane. On exclut alors de la distribution les 10 % des valeurs les plus faibles et les 10% des valeurs les plus fortes. Il se note (X_{d9}-X_{d1}).

3) Mesures de la dispersion statistique en référence à la moyenne arithmétique

3-1 Ecart absolu moyen : Ce paramètre est la moyenne arithmétique de la valeur absolue des écarts à la moyenne.

$$\textbf{E.A.M. de X} = \sum_{i=1}^{N} |Xi - \bar{X}| / N$$

3-2 Variance et écart-type :

- Variance : La variance, notée $(\sigma_x)^2$ est la moyenne du carré des écarts à la moyenne.

$$(\sigma_x)^2 = \sum_{i=1}^{N} (Xi - \bar{X})^2 / N$$

La variance n'est pas un paramètre de dispersion absolue mais plutôt une mesure globale de la variation d'un caractère de part et d'autre de la moyenne arithmétique (quantité d'information). Pour obtenir un paramètre de dispersion absolue, on effectue la racine carrée de la variance, appelé **écart-type** et que l'on note σx

- Ecart-type : L'écart type, noté σx est la racine carré de la moyenne du carré des écarts à la moyenne, c'est à dire la racine carrée de la variance.

$$\sigma_x = \sqrt{(\sigma_x)^2}$$

L'écart-type est une mesure de dispersion par rapport à la moyenne qui intègre les valeurs algébriques des écarts à la moyenne et qui pourra, à ce titre être réintroduite dans des calculs algébriques ultérieurs. Elle présente de plus l'avantage d'avoir une **signification probabiliste** que ne possède pas l'écart absolu moyen. La théorie des probabilités permet en effet d'estimer la chance qu'a une valeur d'être éloignée de la moyenne de plus d'un certain nombre d'écart-types.

Lorsqu'une distribution est **gaussienne** (on dit aussi "**normale**") les probabilités de trouver les valeurs a une distance donnée de la moyenne sont les suivantes :

68.3 % des valeurs sont comprises entre $(\bar{x}-\sigma_x)$ et $(\bar{x}-\sigma_x)$

95.5 % des valeurs sont comprise entre $(\bar{x}-2\sigma_x)$ et $(\bar{x}+2\sigma_x)$

99.7 % des valeurs sont comprises entre $(\bar{x}-3\sigma_x)$ et $(\bar{x}+3\sigma_x)$

LES PARAMETRES DE DISPERSION RELATIVE

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur. Dans le cas contraire, la comparaison n'est possible qu'en ayant recours à des mesures de **dispersion relative**, c'est à dire *en effectuant le rapport entre un paramètre de dispersion absolue et la valeur centrale qui lui tient de référence* .

Un paramètre de dispersion relative est une mesure de **l'écart relatif** des valeurs d'une distribution à une valeur centrale. C'est donc le rapport d'un paramètre de dispersion absolue divisé par une valeur centrale. On obtient un nombre sans dimension qui peut être exprimé en %.

Dispersion relative = Paramètre de dispersion absolue/Valeur centrale _____

- le coefficient interquartile relatif

$= (X_{q3}-X_{q1})/$ médiane X

- l'écart moyen relatif
= E.A.M. / $\overline{X}$

- le coefficient de variation
= σ_x / $\overline{X}$

Remarque très importante : Le calcul d'un paramètre de dispersion relative n'est possible que pour les caractères quantitatifs **positifs** (toutes les modalités sont des nombres positifs).

Vocabulaire 4 Discrétisations d'une distribution statistique et représentations graphiques

4.1 Distribution statistique et discrétisation

Distribution statistique: pour un caractère quantitatif, ensemble ordonné des modalités prises par le caractère X par l'ensemble des éléments de E

- *Le **tableau de dénombrement*** donne un *résumé numérique* d'une distribution statistique.
- *Les représentations graphiques* donnent un *résumé visuel* d'une distribution statistique
- *Les représentations cartographiques* donnent une image de la répartition dans l'espace de la distribution statistique

La construction du tableau de dénombrement et des représentations graphiques sera différente selon que le caractère étudié *quantitatif discret, quantitatif continu, ou caractère qualitatif.*

CARACTERES DISCRETS

*** Tableau de dénombrement :** le tableau de base se compose de trois colonnes:

1- La liste ordonnée des valeurs X_i du caractère, ou la suite des classes de valeur du caractère, si les modalités numériques ont été regroupées en classes. Les modalités d'un caractère qualitatif n'ont pas d'ordre.

2- L'effectif n_i d'une valeur (ou classe de valeurs) de X_i est le nombre d'éléments qui prennent cette valeur (ou classe de valeurs) dans la distribution observée. La somme du nombre des éléments concernés par chaque modalité (caractère qualitatif) ou par chaque valeur (caractère quantitatif) donne le nombre d'éléments N de l'ensemble observé. Pour k modalités on a :

$$\sum_{i=1}^{K} n_i = N$$

La fréquence simple f_i d'une modalité ou d'une valeur X_i est le rapport entre l'effectif correspondant à cette modalité et l'effectif total. 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%.

$f_i = n_i / N$ (fréquence sous forme décimale)
$f_i = 100 * f_i / N$ (fréquence en pourcentage)

La somme des fréquences simples est égale à 1 (ou à 100 %) des éléments.

$$\sum_{i=1}^{K} f_i = 1 = 100\%$$

CARACTERES CONTINUS

*** Tableau de dénombrement : Il est un tableau qui a autant de lignes que l'on a retenu de classes de valeurs quand on a discrétisé la variable**

1- Discrétisation de la variable ou mise en classes. Les classes correspondent à une partition de l'ensemble de l'intervalle de variation du caractère. (intervalle allant de la valeur minimum prise par X dans l'ensemble étudié, à la valeur minimum prise par X dans l'ensemble étudié). Ces classes doivent être 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. Chacune des k classes, j étant une classe quelconque, est définie par une borne inférieure X_{Binfj} et une borne supérieure X_{Bsupj}, n_j est le nombre d'éléments compris dans l'intervalle $[X_{Binfj}; X_{Bsupj}[$

2- L'amplitude de la classe :

Pour chaque classe j, $A_j = X_{Bsup\,j} - X_{Binf\,j}$

3- Le centre de la classe :

$C_j = [\, X_{Bsupj} + X_{Binf\,j}] / 2$

4- L'effectif de la classe A chaque classe j correspond un effectif n_j qui est le nombre d'éléments de l'ensemble E concentrés dans cette classe

5- La fréquence simple de la classe :

$f_i = n_j / N$ ou $100^* n_j / N$

6- La fréquence moyenne ou densité d'effectif :

$f_{mj} = f_j / A_j$

7- La fréquence cumulée ascendante de la classe j est la proportion d'éléments qui pour le caractère X enregistrent une valeur *inférieure* à celle de sa borne supérieure.

$$F_{asc\,j} = f_j + \sum_{j=1}^{j=j-1} f_j$$

8- La fréquence cumulée descendante de la classe j est la proportion d'éléments qui, pour le caractère X enregistrent une valeur *supérieure* à celle de sa borne inférieure

$F_{desj} = 1 - F_{ascj}$

4.2 Représentations graphiques d'une distribution statistique

CARACTERES QUALITATFS

Diagramme en bâtons *:* Il est établi à partir d'un tableau de dénombrement. Il présente en abscisse la suite des modalités du caractère X (pas d'orientation donnée à l'abscisse), et en ordonnée l'effectif ou la fréquence de chacune des modalités.

CARACTERES QUANTITATIFS DISCRETS

Diagramme en barres : Il est établi à partir d'un tableau de dénombrement. C'est la représentation graphique normale d'un caractère discret. Il présente, en abscisse la suite ordonnée des valeurs du caractère X, et en ordonnée leur fréquence simple ou leur effectif. Les bâtons ne doivent pas être jointifs car le caractère est discret.

CARACTERES QUANTITATIFS CONTINUS

1- ***L'histogramme*** est établi à partir d'un tableau de dénombrement. C'est une représentation bi-dimensionnelle qui présente chaque classe sous la forme d'une rectangle dont la **surface est proportionnelle à l'effectif ou à la fréquence de la classe**. Pour ce faire la base (sur l'axe Ox) est proportionnelle à l'**amplitude** de la classe et la hauteur (sur l'axe Oy) est proportionnelle à la **fréquence moyenne** de la classe.:

Surface du rectangle = (longueur x largeur) = (f_{mj} x A_j) = (f_j / A_j) x A_j = f_j

La **légende** d'un histogramme est constituée par un élément de surface dont on indique la part des éléments qu'il représente (exprimée en fréquence simple ou en effectif).

L' histogramme permet une double lecture de la distribution statistique:

- La hauteur du rectangle (ordonnée) renseigne sur la densité des éléments dans chaque classe. Une forte fréquence moyenne indique une concentration des valeurs, une faible fréquence moyenne indique une dispersion des valeurs.
- La surface des rectangles renseigne sur l'effectif de chaque classe.

2- Les ***courbe des fréquences cumulées*** ascendantes et descendantes sont établies à partir du tableau de distribution statistique ou à partir d'un tableau de dénombrement. C'est un graphique bi-dimensionnel représentant en abscisse les modalités j du caractère X ou les bornes des classes (dans le cas d'une variable mise en classes) et en ordonnée, les fréquences cumulées. Sa construction est la suivante :

- La fréquence cumulée ascendante On porte en abscisse les valeurs correspondant aux bornes des classes de la partition du caractère X et en ordonnée les fréquences cumulées ascendantes correspondantes. On repère les points correspondant à ces couples de valeurs auxquels on ajoute le point qui a pour abscisse la borne inférieure de la première classe, qui correspond à X_{min}, et pour ordonnée 0 (en effet, 0 ou 0% des éléments enregistrent des valeurs inférieures à X_{min}). On procède à

une extrapolation linéaire entre ces points (hypothèse d'équirépartition des éléments dans chaque classe), en reliant ces points par une droite.

- La fréquence cumulée descendante On porte en abscisse les valeurs correspondant aux bornes des classes de la partition du caractère X et en ordonnée les fréquences cumulées descendantes correspondantes. On repère les points correspondant à ces couples de valeurs auxquels on ajoute le point qui a pour abscisse la borne supérieure de la dernière classe, qui correspond à X_{max}, et dont l'ordonnée est 0 (en effet, 0 ou 0% des éléments enregistrent des valeurs supérieures à X_{max}). On procède à une interpolation linéaire entre ces points (hypothèse d'équirépartition des éléments dans chaque classe), en les reliant par une droite.

Les courbes obtenues se croisent au point de fréquence cumulée ascendante ou descendante 0.5 ou 50%. La valeur du caractère X correspondant à cette fréquence cumulée 0.5 ou 50% et que l'on peut directement lire sur le graphique en ordonnée est la **médiane** (50 % des éléments sont supérieurs à cette valeur de X et 50 % lui sont inférieurs). On notera ce point $X_{méd}$

5. Comment transformer l'unité de mesure d'un caractère, ou éliminer l'effet de cette unité

La comparaison de plusieurs caractères quantitatifs ou bien, celle de caractères quantitatifs et qualitatifs,, n'est généralement pas possible à partir du tableau élémentaire de départ, car les caractères à comparer peuvent avoir :

- des unités de mesure différentes.
- des ordres de grandeur différents (valeurs centrales)
- des dispersions différentes (paramètres de dispersion).
- des modalités qualitatives

Il est donc généralement nécessaire de transformer les caractères quantitatifs à étudier avant de pouvoir les comparer les uns aux autres. Quatre types de transformations sont proposées :

1) La discrétisation, ou transformation des valeurs d'une distribution en modalités d'un caractère qualitatif nominal

*** dichotomie (2 classes)**

On fixe un seuil X_s qui définit la limite entre les valeurs fortes et faibles de X_i. X_s peut être une valeur centrale (moyenne, médiane) ou bien une valeur qui possède une signification particulière pour l'interprétation. On crée la variable qualitative X' ayant deux modalités :

X_s peut être une modalité quelconque de X

*** élargissement :** au lieu de se limiter à deux classes, on peut décider que chaque variable se composera n classes ("fort/moyen/faible", très grand, grand, petit très petit, etc..).

si X_s = X médiane, alors la *médiane est une limite de classe* et on construit des classes *d'effectifs égaux* de part et d'autre de la médiane

si X_s = X moyenne, alors la *moyenne est une limite de classe* et on construit des classes d'amplitude égale de part et d'autre de la moyenne.

Si on souhaite comparer deux caractères, la discrétisation doit être effectuée dans les deux cas en suivant strictement le même principe.

2) La transformation d'un caractère quantitatif en un caractère ordinal

Chaque caractère quantitatif X_i est transformé en un caractère qualitatif ordinal X'_i qui indique le rang pris par i dans la série X.

X'_i => rang de X_i dans la distribution statistique de X

La méthode ne pose pas de problème mais il faut faire attention à deux choses

- l'ordre de classement (croissant ou décroissant) doit être spécifié et être le même pour les deux caractères
- lorsqu'il y a des ex-aequo, on leur attribue comme rang la moyenne des places qu'ils auraient occupées s'ils avaient été à la suite les uns des autres. On reprend ensuite le classement après les rangs virtuellement occupés.

exemple rang des régions françaises pour la densité de population en 1946 et en 1990

3) La standardisation

La standardisation est la transformation la plus efficace quand on veut comparer deux variables quantitatives. Elle consiste à opérer une double transformation de **centrage** et de **réduction**.

Centrage : L'opération de centrage consiste à transformer un caractère X en un caractère X' qui exprime les écarts positifs ou négatifs par rapport à une valeur de référence qui est la moyenne arithmétique de la distribution

$$\mathbf{X'_i = (} X_i - \overline{X} \mathbf{)}$$

Réduction : l'opération de réduction consiste à transformer une variable X en la divisant par l'écart-type de la distribution

$$\mathbf{X'_i = X_i / \sigma_X}$$

Dans la plupart des cas, on utilise l'écart-type pour effectuer la réduction.

Standardisation: une variable standardisée (on dit aussi **centrée-réduite**) a été centrée par la moyenne et réduite par l'écart-type :

$$\mathbf{X'_i = (} X_i - \overline{X} \mathbf{)/\sigma_X}$$

Une variable standardisée (centrée-réduite) possède une moyenne de 0 et un écart- type de 1. Elle exprime l'écart d'un élément de la distribution à la moyenne, mesuré en écarts-types. L'unité de mesure de la variable d'origine a donc disparu et il est toujours possible de comparer deux variables standardisées.

Transformer les caractères pour pouvoir les comparer 2 à 2

Types de caractères à comparer	**quantitatif**	**ordinal**	**qualitatif nominal**
quantitatif	comparaison après élimination de l'effet de l'unité de mesure (standardisation)	après transformation du caractère quantitatif en caractère ordinal	après transformation du caractère quantitatif en caractère qualitatif (discrétisation)
ordinal	après transformation du caractère quantitatif en caractère ordinal	comparaison directe	après transformation du caractère ordinal en caractère qualitatif
qualitatif nominal	après transformation du caractère quantitatif en caractère qualitatif (discrétisation)	après transformation du caractère ordinal en caractère qualitatif	comparaison sous conditions