CHAPITRE 1 : Distribution statistique à une dimension

Section 1 : Vocabulaire élémentaire de la statistique descriptive

1. Population et individu

Définition

→ On appelle population statistique, tout ensemble d'unités statistiques constituant les unités observées.

→ On appelle individu (ou unité statistique), tout élément de la population étudiée.

Remarque

La détermination avec précision de la population et des individus qui la composent conditionne l'homogénéité des unités observée et la fiabilité des résultats.

2. Variable statistique

2.1. Définition

On appelle variable statistique (ou caractère), une application (relation) qui associe à chaque individu de la population, une observation particulière.

2.2. Type d'une variable statistique

a) Variable qualitative

Une variable est qualitative si elle est liée à un ensemble d'observations non mesurables.

Exemple : La population active tunisienne peut être caractérisée par :

- Le sexe (masculin ou féminin)
- La catégorie professionnelle (cadres, employés, ouvriers, etc...)

La nature qualitative d'une variable s'exprime donc par l'appartenance à une catégorie ou rubrique d'un ensemble fini.

b) Variable quantitative

Une variable est quantitative si l'ensemble des observations est un ensemble de nombres. Ces observations expriment donc des valeurs numériques (quantitatif = mesurable).

Exemple : La Catégorie Hôtelière, La Capacité en lits,…

Les variables quantitatives peuvent être discrètes ou continues :

→ Une variable quantitative est dite discrète lorsqu'elle prend un nombre fini ou dénombrable de valeurs (La Catégorie Hôtelière).

La variable «nombre d'enfants par ménage» est une variable quantitative discrète.

→ Une variable quantitative est dite continue, lorsqu'elle prend toutes les valeurs d'un intervalle réel.

La variable «La Capacité en lits» peut être envisage comme une variable quantitative continue. Ses valeurs sont le plus souvent regroupées par intervalle.

3. Effectifs

Définition

On appelle effectif d'une valeur (ou rubrique) donnée x_i le nombre de fois ou cette valeur (ou rubrique), apparait dans la population statistique étudiée. Ce nombre est note n_i. L'effectif est parfois appelé fréquence absolue.

On appelle effectif total de la population étudiée, noté n, la somme des p effectifs particuliers n_i correspondant à chacune des valeurs (ou rubriques), soit :

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

Le symbole $\sum$ (lu «somme») permet une écriture synthétique de la somme des p effectifs $n_1, n_2 \ldots n_p$. On lit alors « n = somme des n_i (pour i variant de 1 à p) ».

4. Fréquences

On appelle fréquence de la valeur (ou modalité) x_i, notée f_i, le rapport de l'effectif n_i correspondant a la valeur x_i et de l'effectif n de la population observée.

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

Ce rapport est égal au pourcentage d'individus présentant la valeur (ou modalité) x_i par rapport à l'ensemble de la population observée. f_i est toujours comprise entre 0 et 1.

Pour une série statistique présentant p valeurs (ou modalités), on a :

$$\sum_{i=1}^{p} f_i = f_1 + f_2 + \ldots\ldots\ldots + f_p = 1$$

Remarque :

→ Parfois on peut rencontrer le terme de fréquence relative pour les fréquences.

→ L'emploi des fréquences ou fréquences relatives s'avère utile pour comparer deux distributions de fréquences établies à partir d'échantillons de taille différente.

On appelle fréquences cumulées ou fréquences relatives cumulées en x_i, le nombre f_i cum tel que $F_j = f_i cum = \sum_{p=1}^{i} f_p$

Section 2 : Représentation des données

Il existe plusieurs niveaux de description statistique : la présentation brute des données, des présentations par tableaux numériques, des représentations graphiques et des résumés numériques fournis par un petit nombre de paramètres caractéristiques.

1. Séries statistiques

Une série statistique correspond aux différentes modalités d'un caractère sur un échantillon d'individus appartenant à une population donnée.

Le nombre d'individus qui constituent l'échantillon étudié s'appelle la taille de l'échantillon.

2. Tableaux statistiques

Le tableau de distribution de fréquences est un mode synthétique de présentation des données. Sa constitution est immédiate dans le cas d'un caractère discret mais nécessite en revanche une transformation des données dans le cas d'un caractère continu.

2.1 Caractère qualitatif

Modalité *Numéro i*	*Effectif* n_i	*Fréquence* f_i
1 *2* …….. *p*	n_1 n_2 …….. n_p	f_1 f_2 …….. f_p

On a les relations suivantes:

$$n = \sum_{i=1}^{p} n_i \qquad f_i = \frac{n_i}{n} \qquad \sum_{i=1}^{p} f_i = 1$$

Remarque : les fréquences peuvent être exprimées dans le tableau en pourcentage, dans ce cas : $\sum_{i=1}^{p} f_i = 100$

2.2 Caractère quantitatif

a) Caractère quantitatif discret

Valeurs observées x_i	*Effectifs* n_i	*Fréquences* f_i	*Fréquences cumulées croissantes* F_i
x_1 x_2 …….. x_P	n_1 n_2 …….. n_P	f_1 f_2 …….. f_P	F_1 F_2 …….. F_P

b) Caractère quantitatif continu

Classes numéro i $[b_i ; b_{i+1}[$	*Centres* c_i	*Effectifs* n_i	*Fréquences* f_i	*Fréquences cumulées croissantes* F_i
$[b_1 ; b_2[$	c_1	n_1	f_1	F_1
$[b_2 ; b_3[$	c_2	n_2	f_2	F_2
…….	……	……	……	……
$[b_P ; b_{P+1}[$	c_P	n_P	f_P	F_P

Remarque :

- Une classe est un intervalle fermé à gauche et ouvert à droite, du type $[b_i ; b_{i+1}[$.

- Le centre d'une classe est $c_i = \frac{b_i + b_{i+1}}{2}$

- L'amplitude d'une classe est $a_i = b_{i+1} - b_i$

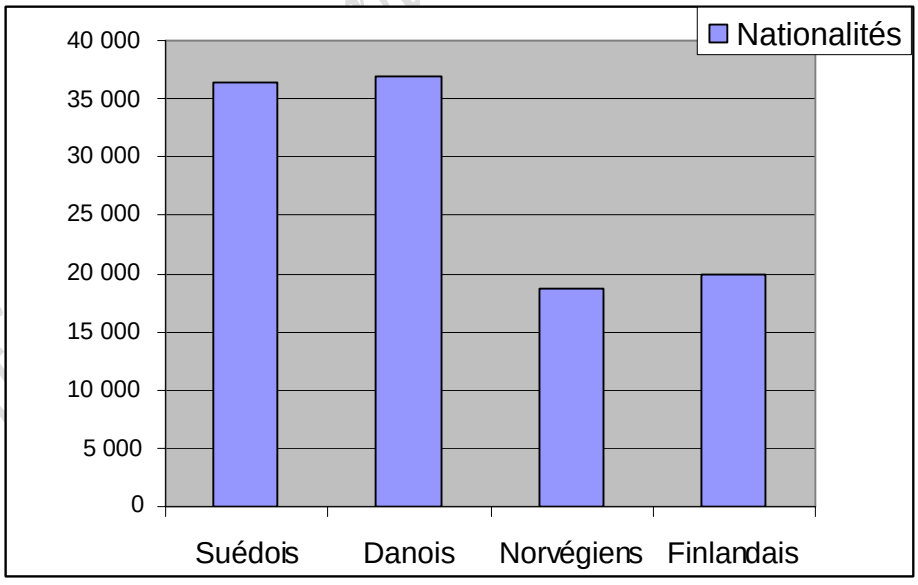
3. Représentations graphiques

3.1 Caractères qualitatifs

a) Diagrammes à bandes

On appelle diagramme à bandes un graphique qui, à chaque modalité de la variable qualitative associe un rectangle de base constante dont la hauteur est proportionnelle à l'effectif.

Figure 1-1 : Diagramme à bandes (verticales) :
nombre d'arrivée aux frontières des scandinaves par nationalité pour l'année 2005

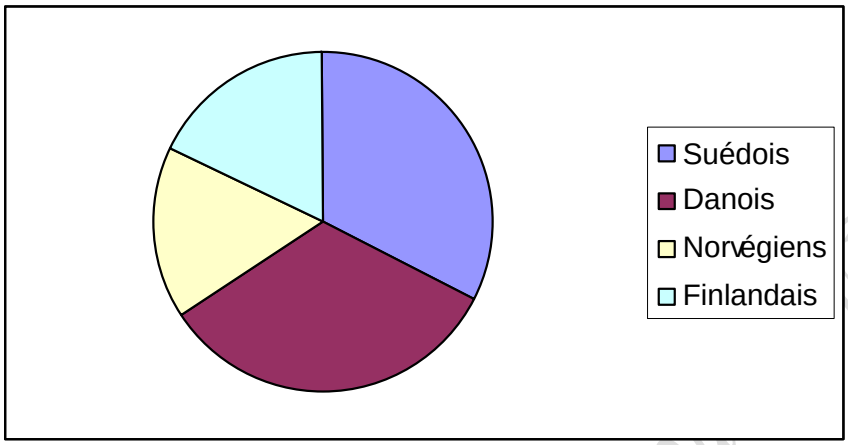


b) Diagrammes à secteurs

On appelle diagramme à secteurs un graphique qui divise un disque en secteurs angulaires dont les angles au centre sont proportionnels aux effectifs de chaque modalité.

Pour une modalité donnée M_i, d'effectif n_i, l'angle au centre α_i, correspondant est donné (en degré) par : $\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360$

Figure 1-2 : Diagramme à secteurs :
nombre d'arrivée aux frontières des scandinaves par nationalité pour l'année 2005



3.2 Caractères quantitatifs

3.2.1 Caractère quantitatif Discret

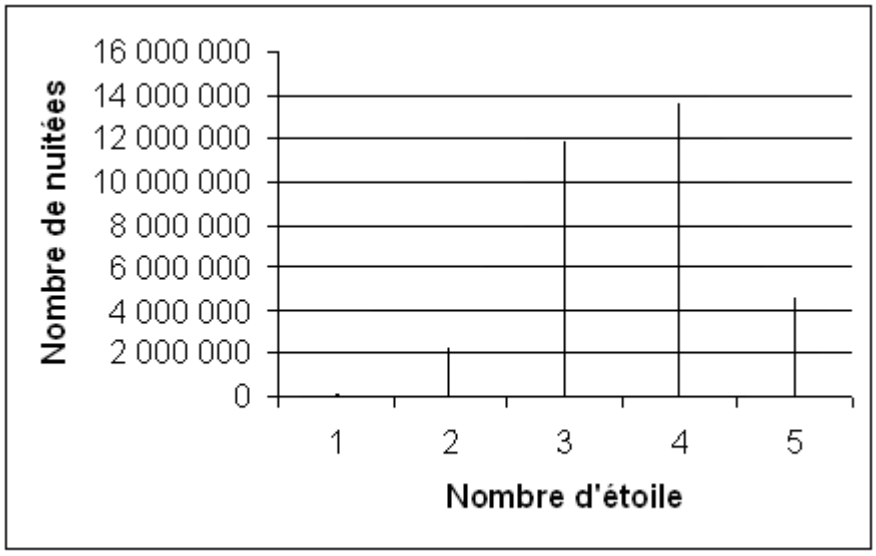
a) Diagramme en bâtons

On appelle diagramme en bâtons un graphique qui associe à chaque valeur de la variable un segment (bâton) dont la hauteur est proportionnelle à l'effectif.

Remarque

On suppose les valeurs observées de la variable quantitative discrète, ordonnées par ordre croissant.

Figure 1-3 : Diagramme en bâtons :
Nombre de nuitées des non résidants par nombre d'étoile en 2005



b) Diagrammes cumulatifs

On appelle diagramme cumulatif, la courbe représentative de la fonction de répartition.

On appelle fonction de répartition d'un caractère X, l'application notée F, dont l'ensemble de départ est R et l'ensemble d'arrivée, l'intervalle $[0,1]$.

$F : R \rightarrow [0,1]$

$X \rightarrow F(x) = P(X<x)$

$F(x)$ = proportion des individus dont la valeur du caractère est strictement inferieure à x.

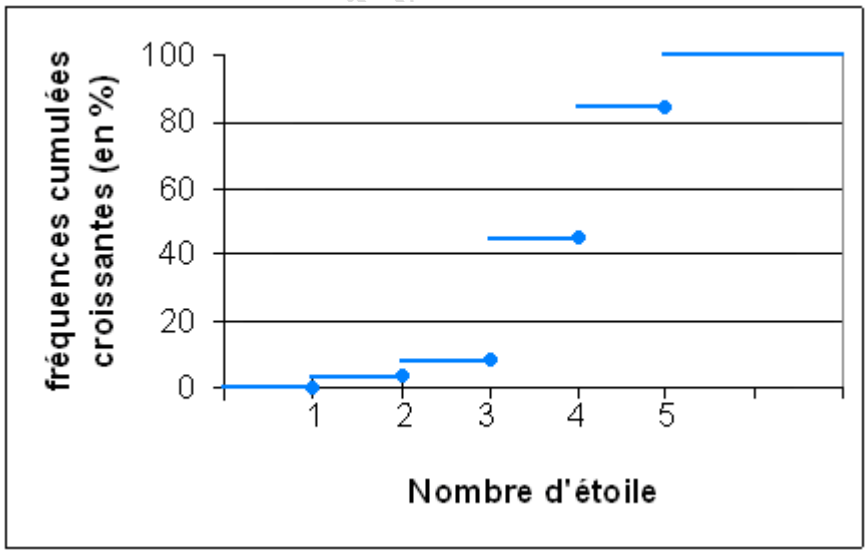
$F(x) = fréquence$ de $(X < x) = f_1 + f_2 + ... + f_p$

Remarque

Le plus souvent, le diagramme cumulatif est obtenu à partir des fréquences cumulées croissantes. Dans le cas d'une variable discrète, la courbe cumulative se présente comme une courbe en escalier (La fonction F est dans le cas discret, une fonction constante par intervalle) Chaque segment de cette courbe en escalier est ouvert à gauche et ferme droite (sauf le dernier).

Si on définit la fonction de répartition par $F(x) = P(X \leq x)$, alors les segments deviennent fermés à gauche et ouverts à droit (sauf le premier).

Figure 1-4 : Courbe en escalier du
Nombre de nuitées des non résidants par nombre d'étoile en 2005



3.2.2 Caractères quantitatifs continus

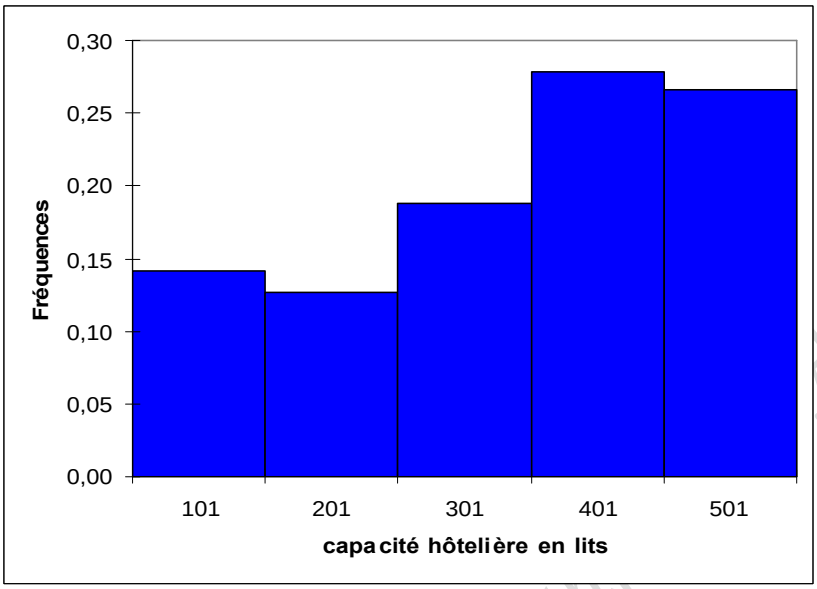
a) Histogrammes

On appelle histogramme un diagramme comparé d'un ensemble de rectangles contigus (adjacents), chaque rectangle, associé à chaque classe, ayant une surface proportionnelle à l'effectif (ou à la fréquence) de cette classe. Deux cas peuvent se présenter :

→ *Classes d'amplitudes égales* :

Lorsque les classes sont d'amplitudes égales ($a_i = a$)

Figure 1-5 : Histogramme (amplitude égales) :
Capacité Hôtelière en lits pour la région de Nabeul-Hammamet



→ *Classes d'amplitudes inégales :*

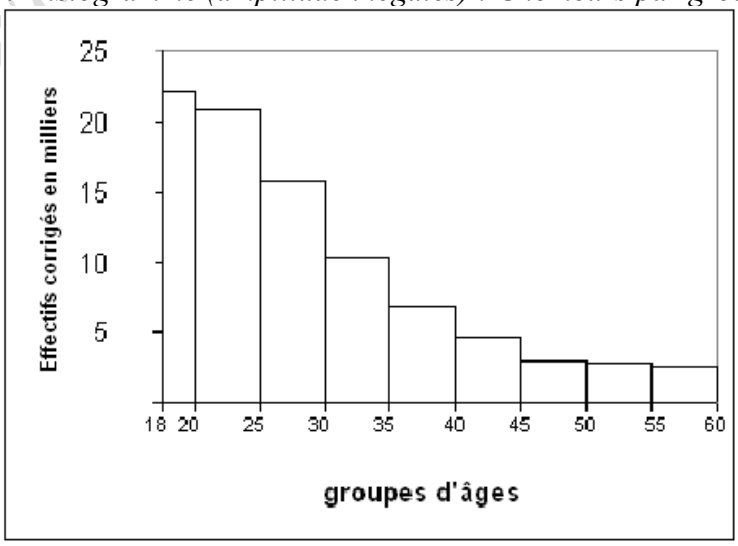
Lorsqu'au moins deux classes ont des amplitudes différentes, la hauteur proportionnelle à l'effectif ne permet plus de construire un histogramme. En effet, la surface de chaque rectangle n'est plus proportionnelle à l'effectif (conformément à la définition).

D'où la nécessité de corrigé les fréquences (ou les effectifs).

$$f_i^c = \frac{f_i}{a_i} \qquad \text{ou} \qquad n_i^c = \frac{n_i}{a_i}$$

Avec: f_i^c : fréquence corrigé ; n_i^c : effectif corrigé

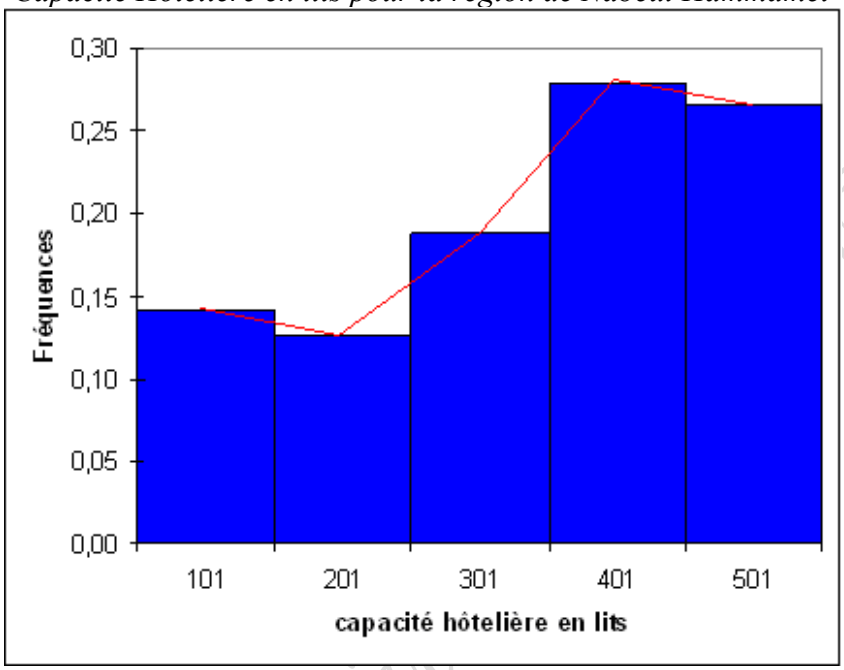
Figure 1-6: Histogramme (amplitude inégales) : Chômeurs par groupe d'âges



b) Polygone des fréquences

On obtient un polygone de fréquences en joignant les milieux des segments supérieurs de chaque rectangle de l'histogramme (l'aire du polygone des fréquences est l'aire de la surface par la ligne polygonale et l'axe des abscisses).

Figure 1-7 : Polygone des fréquences (amplitude égales) : Capacité Hôtelière en lits pour la région de Nabeul-Hammamet



c) Courbe des fréquences cumulées croissantes

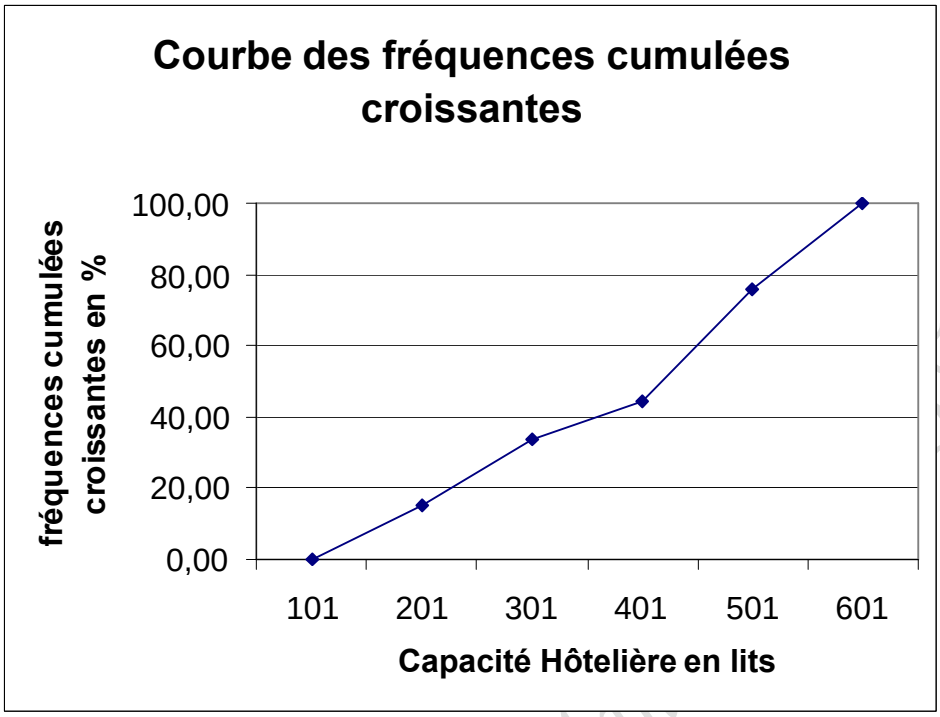
Définition

On appelle courbe des fréquences cumulées croissantes la représentation graphique de la fonction de répartition de la variable x.

Les données étant groupées en classes, la fréquence cumulée croissante F_i associée à la classe $n° : i$ représente la proportion d'individus de la population pour lequel la variable prend une valeur inferieure (strictement) à la limite supérieure b_i de la classe $n° : i$.

En pratique, la courbe des fréquences cumulées croissantes est obtenue en joignant, dans un système d'axes orthogonaux, les points d'abscisse b_i (extrémité de la classe $n° : i$) et d'ordonnée F_i (fréquence cumulée croissante correspondante) [Remarque : Joindre les points de coordonnées (b_i, F_i) par des segments revient à faire l'hypothèse d'une répartition uniforme des individus à l'intérieur des classes].

Figure 1-8 : Courbe des fréquences cumulées croissantes : Capacité Hôtelière en lits pour la région de Sousse-Kairouan



3.3 Graphiques spécialises

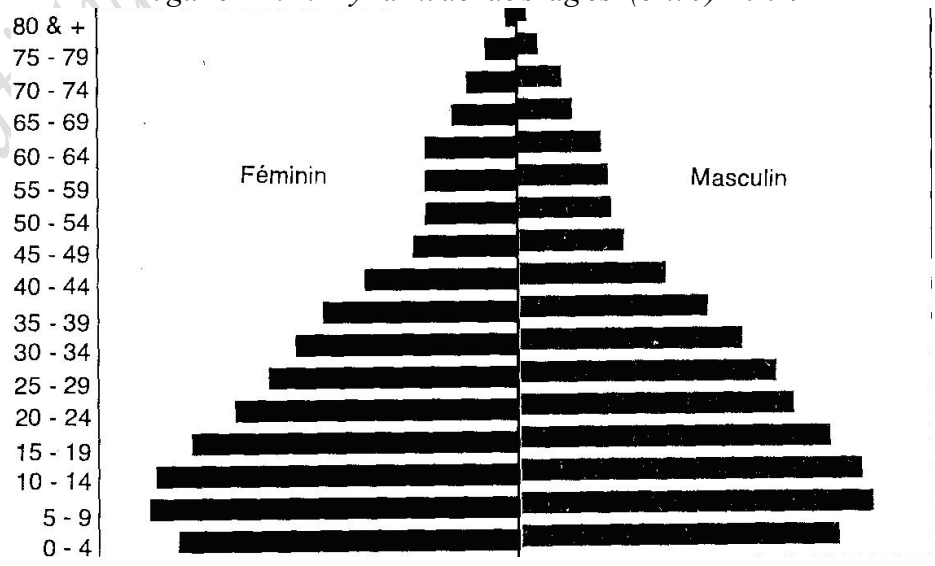
a) Pyramide des âges

Exemple : Pyramide des âges en Tunisie en 1994 :

La pyramide est une version particulière de l'histogramme, utilisée pour repartir une population en fonction de l'âge (variable quantitative continue) et du sexe des individus (variable qualitative secondaire).

Le graphique se présente comme un histogramme double, les valeurs de la variable principale (l'âge) étant portées sur l'axe vertical.

Figure 1-9 : Pyramide des âges (en%) 1994

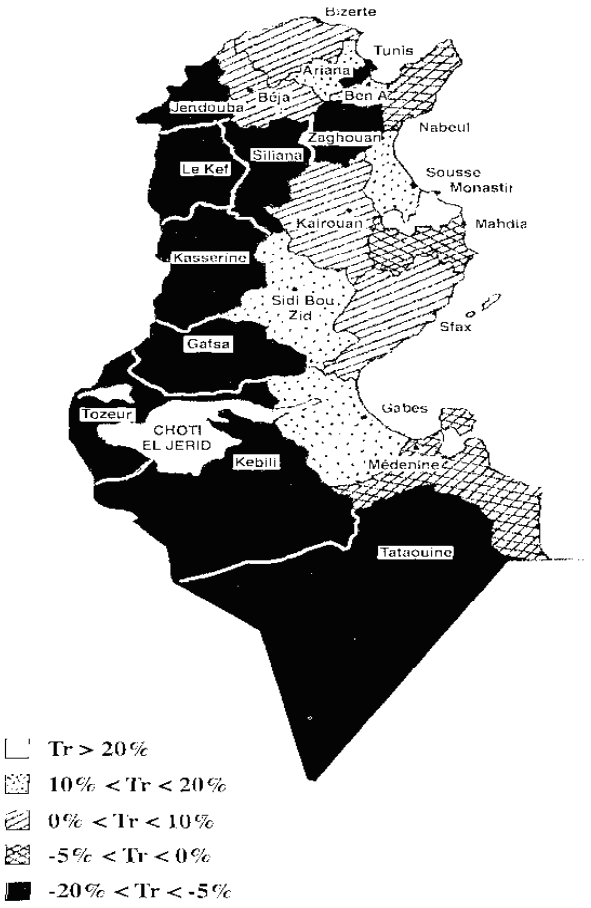


b) Diagramme cartographique (ou cartogramme)

Lorsqu'une étude statistique concerne des données géographiques, les individus ou les modalités de la variable étudiée sont des unités spatiales (des régions par exemple).
On petit alors utiliser une carte pour visualiser les valeurs associées a un caractère et a une unité spatiale.

La population étudiée est « l'ensemble des gouvernorats de la Tunisie ». La variable étudiée est « l'écart global en taux (L'écart global en taux, représente ici le différentiel de croissance de l'emploi un gouvernorat donné et la nation toute entière sur la période 1984-1989. Un écart global positif (négatif), signifie que le taux de croissance de l'emploi du gouvernorat considéré est supérieur (inferieur), au taux do croissance moyen au niveau national)». Elle est quantitative continue. Une décomposition des valeurs de la variable en 5 classes d'inégale amplitude conduit à la figure suivante :

Figure 1-10 : Diagramme Cartographique
Ecart global en taux (Tr) par gouvernorat (1984-1989)



c) Diagramme polaire

Ce diagramme (*figure 1-10*) est utilise lorsqu'on cherche à comparer, pour une variable donnée, des observations relatives à plusieurs sous populations (d'une même population). On suppose que la variable (qualitative ou quantitative) étudiée présente *p* modalités. Le plan est divisé en *p* secteurs délimités par des axes associes aux différentes modalités de la variable. On reporte sur chaque axe la valeur correspondant à la modalité concernée ; la distance du point à l'origine est proportionnelle à la valeur correspondante. Les points correspondants à une même sous population peuvent être joints (*figure 1-10*).

Les données du tableau précisent le nombre de chômeurs selon la durée du chômage (en mois) et le niveau d'instruction enregistres en 1997(Source : enquête national sur l'emploi, 1999).

Tableau: Nombre de chômeurs en Tunisie, enregistrés en 1997

Durée de chômage (en mois)	néant	primaire	secondaire	supérieur	total
[0,1[	1880	5089	2453	138	9560
[1,3[	12616	39895	16193	731	69435
[3,6[	12108	43318	16346	1072	72844
[6,9[	4002	19797	13263	1657	38719
[9,12[	2112	8419	8582	1816	20929
[12,24[	16224	93402	83346	11708	204680

Figure 1-10 : Diagramme Polaire