TD

JEAN-LOUIS MONINO

STATISTIQUE DESCRIPTIVE

- QCM et exercices corrigés
- 4 sujets d'examen corrigés
- Avec rappels de cours

5e ÉDITION

DUNOD

Ressources numériques

En complément de cet ouvrage, vous trouverez 2 sujets d'examen corrigés en ligne.

http://www.dunod.com/contenus-complementaires/9782100758777

Mise en page : Belle page

11, rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-075877-7

Sommaire

Avant-propos

Le monde est devenu numérique et les avancées technologiques ont démultiplié les circuits d'accès aux données, à leur traitement et à leur diffusion. Les données sont accessibles à tous et partout sur la planète. Le nombre d'internautes en 2014 était de 2,9 milliards soit 41 % de la population mondiale. Le besoin de connaissance se fait sentir afin d'appréhender cette multitude de données. Il faut renseigner, informer et former en masse. L'essor des technologies connexes, comme l'avènement de l'internet, des réseaux sociaux, des « cloud-computing » (usines numériques), a accru les volumes disponibles de données (Data). Actuellement, chaque individu crée, consomme, utilise de l'information numérique : plus de 3,4 millions d'e-mails envoyés dans le monde chaque seconde, soit 107 000 milliards annuels avec 14 600 mails par an et par personne, mais plus de 70 % sont des spams. Des milliards de contenus sont partagés sur les réseaux sociaux comme par exemple sur Facebook, plus de 2,46 millions chaque minute. La statistique est omniprésente, tous les jours, à travers les médias notamment, chacun d'entre nous se trouve exposé à une foule d'informations issues d'études quantitatives : sondages d'opinion, « baromètres » de popularité des hommes politiques, enquêtes de satisfaction, indicateurs de la comptabilité nationale,…

Les tableurs permettent d'effectuer des calculs parfois complexes en des temps très faibles ; aussi leur parfaite maîtrise nous paraît intimement liée aux objectifs de l'enseignement de la statistique descriptive. Les méthodes statistiques, notamment les plus sommaires telles que celles liées à la représentation graphique des données numériques, sont souvent choisies sans véritable connaissance de leur utilité et des limites de leur utilisation. Le développement de l'informatique a aussi pour conséquence de faire évoluer l'enseignement de la statistique. Il est à l'heure actuelle inconcevable de n'illustrer un cours de statistique qu'à l'aide de calculs effectués « à la main ». Pourtant, la plupart des ouvrages dédiés aux étudiants sont conçus pour s'appliquer à un enseignement magistral de la statistique descriptive.

Tous ces constats nous ont amenés à proposer un ouvrage « charnière », s'inscrivant dans une approche nouvelle de l'enseignement de la statistique descriptive et destiné à un très large public ; un ouvrage permettant d'utiliser les possibilités offertes par les tableurs ou logiciels de statistique.

Compte tenu de l'expérience de l'auteur, enseignants dans les disciplines de sciences économiques et de gestion, *Statistique descriptive* – TD est particulièrement adapté aux étudiants des premiers cycles en licence d'économiques, gestion. Il veut cependant répondre aussi au besoin de tous les publics désireux d'assimiler et d'appliquer à des cas pratiques ou concrets les concepts principaux de la statistique descriptive : étudiants en médecine, pharmacie, psychologie, langues étrangères appliquées, mathématiques appliquées aux sciences sociales, mais aussi étudiants de cycles courts : BTS, DUT, etc. et de classes préparatoires. Les stagiaires de la formation continue y trouveront par ailleurs un outil adapté à leur souci de compréhension rapide et de pragmatisme.

Le lecteur intéressé par des prolongements de cet ouvrage est vivement invité à se connecter sur le site Internet http://www.portices.fr/Cours_Statistique où il trouvera d'autres éléments de cours, des exercices et des examens corrigés.

L'auteur remercie les étudiants qui ont testé les exercices de cet ouvrage durant l'année universitaire 2015-2017 : étudiants de l'UFR d'Économie, de l'École Sécurité Environnement Qualité de l'UFR de Droit et Sciences Politiques de l'Université de Montpellier. Je remercie plus particulièrement les étudiants des promotions de la « prépa » ENS Cachan du lycée Jean Mermoz de l'Académie de Montpellier.

Jean-Louis MONINO, Docteur en Sciences économiques I, Maître de Conférences à l'Université Montpellier I. Économètre, Responsable du C2i au sein de la formation continue de l'Université de Montpellier responsable du « Centre de Certification Numérique International ». Il est membre fondateur du réseau « Cercle K2 » et du Réseau de Recherche sur l'Innovation (RRI), Responsable du Diplôme Universitaire – Traitement de l'Information et Intelligence Économique –. Il enseigne la statistique, les méthodes probabilistes, l'analyse des données, les méthodes de sondage, le data Mining,…dans plusieurs universités françaises et à étrangères (Maroc « École des Sciences de l'Information », Burkina Faso « Fondation du « 2iE », Sénégal « Université Virtuelle du Sénégal »,…).

TD 1 Vocabulaire de base

COURS

L'ESSENTIEL

Le terme statistique a un double sens en français :
– il désigne la totalité des données numériques (par exemple, la consommation des ménages français) ou non numériques (la situation matrimoniale) d'un ensemble ;
– il désigne également un ensemble cohérent de méthodes scientifiques qui permettent de résumer l'ensemble des données numériques ou non numériques (par exemple, le calcul de la consommation moyenne des ménages français).

1 Vocabulaire

1.1 La statistique

La statistique est un ensemble de méthodes scientifiques visant le recueil, l'organisation, la présentation de données, ainsi que la modélisation et la construction de résumés numériques. On parle de statistique descriptive lorsqu'on décrit et analyse un ensemble sans tirer de conclusion, de statistique inductive lorsqu'on tire des conclusions sur une partie d'un ensemble et que l'on tente d'étendre ces conclusions sur tout l'ensemble.

1.2 Population. Échantillon. Individu

Nous réalisons une enquête sur les consommateurs de la région de Nantes. Nous considérons que l'ensemble des consommateurs de la région nantaise est la population concernée par l'enquête.

Un ensemble d'éléments, d'objets ou d'individus, regroupés dans une catégorie bien définie est appelé population ou univers.

Mais prendre l'ensemble des consommateurs (population) pour étudier leur comportement vis-à-vis de leur consommation serait trop onéreux. Ainsi, lorsque la population devient trop grande pour être étudiée et analysée dans sa totalité, on considère une partie de cette population (échantillon). Bien sûr, cette partie doit contenir les propriétés fondamentales de la population (ou univers).

Un échantillon est un sous-ensemble de l'ensemble population. Il doit posséder les propriétés fondamentales de l'ensemble dont il est issu.

Le consommateur de la région de Nantes constitue un élément de la population et à ce titre, s'il est désigné pour répondre à un questionnaire, il peut être défini comme l'unité ou l'élément ou le consommateur de référence. Nous analyserons alors son comportement et sa façon d'exercer sa consommation dans la région.

Un individu, ou unité statistique, est un élément qui appartient à une population ou à un échantillon. Un individu est un élément constitutif d'une population ou d'un échantillon qui permet de donner des informations sur les données statis-tiques recherchées.

Pour recueillir les données statistiques (informations numériques ou non numé-riques) sur une population, on utilise soit les enquêtes exhaustives, soit les enquêtes partielles.

Une enquête est dite exhaustive lorsqu'on s'intéresse à l'ensemble des individus qui composent la population étudiée.

Une enquête est dite partielle lorsqu'on s'intéresse à une partie représentative des individus qui composent la population étudiée.

Une enquête exhaustive est appelée recensement. Toute enquête qui n'est pas exhaustive est partielle. Une enquête partielle est appelée enquête par sondage ou plus simplement sondage ; dans ce cas la représentativité joue un rôle déterminant dans la constitution de l'échantillon.

1.3 Caractères. Modalités

Nous avons réalisé une enquête sur 842 commerces de la région Languedoc-Roussillon en collaboration avec la CCI Languedoc-Roussillon[1]. Les commerces ont répondu à la question « Dans quelles proportions a évolué votre chiffre d'affaires pendant les soldes d'été 2016 par rapport aux soldes d'été

1 http://www.languedoc-roussillon.cci.fr/ecommunautes/a/e-develop/32105

2015 ? ». Nous étudions la répartition de cette évolution selon les trois catégories d'achats[1]. Nous obtenons le tableau ci-dessous :

Tableau 1.1 : **Évolution de votre chiffre d'affaires pendant les soldes d'été 2016**

Évolution du chiffre d'affaires	Divers	Équipement de la personne	Équipement du foyer	Total général
Évolution négative de > –20 %	11	41	5	57
Évolution négative de –10 à –20 %	20	61	11	92
Évolution négative de –5 à –10 %	15	74	6	95
Évolution négative de 0 à –5 %	34	123	14	171
Stabilité, pas d'évolution sensible	37	111	8	156
Évolution positive de 0 à 5 %	31	96	10	137
Évolution positive de 5 à 10 %	14	68	3	85
Évolution positive de 10 à 20 %	6	28	2	36
Évolution positive >20 %	0	9	4	13
Total général	168	611	63	842

Ainsi, « Évolution du chiffre d'affaires » est une des multiples questions que nous avons posées aux différents commerces retenus pour l'étude. L'« Évolution du chiffre d'affaires » correspond à un caractère.

Parmi les 842 commerces interrogés, 123 ont répondu avoir eu évolution négative comprise entre 0 à -5 % pour la catégorie « Équipement de la personne ».

Un caractère est une des multiples facettes selon laquelle on peut étudier un individu d'une population ou d'un échantillon. Un individu extrait d'une population ou d'un échantillon peut être étudié selon plusieurs caractères.

Les modalités sont les différentes situations possibles d'un caractère. Une modalité est donc une des réponses possibles à une question.

1 L'équipement de la personne regroupe, dans la nomenclature commerciale, l'ensemble des produits permettant de fournir à une personne une parure. Il comprend les vêtements, chaussures, chapeaux, montres, etc. On l'oppose généralement à l'équipement de la maison, qui comporte des produits qui ne sont pas destinés à être sortis du foyer par l'utilisateur, notamment l'ameublement, l'électroménager, les arts de la table ou le linge de maison. (Source : Wikipédia)

COURS

1.4 Tableaux individus-caractères

Une base de données est un document sur lequel on inscrit l'ensemble des informations utiles à l'étude d'un phénomène. Dans notre exemple sur les consommateurs nantais, nous pouvons dresser une table ou base de données qui comporte 9 601 lignes et autant de colonnes qu'il existe de caractères utiles à l'enquête.

Tableau 1.2 : **Extrait d'une base de données ou tableau individus-caractères**

Numéro	Sexe	Âge	Type d'occupation	Nombre d'actifs	Nombre de voitures
1	Homme	25	Locataire	3	0
2	Homme	35	Propriétaire	4	2
3	Femme	26	Locataire	3	1
4	Homme	38	Propriétaire	1	2
5	Femme	34	Locataire	3	2
6	Homme	33	Locataire	1	3
7	Homme	41	Propriétaire	3	1
8	Homme	64	Locataire	3	3
9	Homme	52	Propriétaire	1	2
10	Femme	39	Propriétaire	3	2
11	Homme	34	Propriétaire	3	2
12	Femme	27	Locataire	4	1
13	Femme	31	Propriétaire	4	1
14	Homme	38	Propriétaire	2	1
15	Femme	64	Propriétaire	2	1
16	Homme	66	Locataire	2	1
17	Femme	71	Propriétaire	2	1
18	Homme	70	Propriétaire	2	1
19	Femme	76	Locataire	1	0
20	Femme	25	Propriétaire	3	0

Un tableau individus-caractères est un recueil de données statistiques. C'est un document où l'on inscrit pour chaque individu d'une population ou d'un échantillon les modalités de réponses à un ensemble de caractères retenus pour l'étude.

1.4 Caractères qualitatifs et quantitatifs

Dans le tableau 1.2, nous avons le caractère « nombre d'actifs » dans le ménage, qui correspond à un caractère quantitatif et le caractère « type d'occupation » qui correspond à un caractère qualitatif, être propriétaire ou locataire.

Lorsque les modalités d'un caractère ne sont pas mesurables, alors le caractère est dit qualitatif.

Lorsque les modalités d'un caractère sont mesurables, alors le caractère est dit quantitatif.

2 Opérateurs somme et produit

Afin de rendre plus compactes les écritures des résumés numériques, nous utilisons des opérateurs mathématiques. L'opérateur *sigma* sert à contracter l'écriture d'une somme de nombres. Il se note $\sum$ et prend la forme :

$$x_1 + \ldots + x_i + \ldots + x_n = \sum_{i=1}^{i=n} x_i$$

Il faut lire de bas en haut : « somme de i égal 1, à i égal n, de x_i ».

Afin de rendre plus compactes les écritures des produits, nous utilisons également un opérateur mathématique. L'opérateur *pi* sert à contracter l'écriture d'un produit de nombres. Il se note $\prod$ et prend la forme :

$$x_1 \cdot \ldots \cdot x_i \cdot x_n = \prod_{i=1}^{i=n} x_i$$

Il faut lire de bas en haut : « produit de i égal 1, à i égal n, de x_i ».

Remarque 1 : se lit (de bas en haut) : « somme de i égale 1, à i égale n, de x indice i »

Remarque 2 : les variables et les indices sont dits muets :

$$\sum_{j=1}^{j=n} x_j = \sum_{i=1}^{i=n} x_i = \sum_{k=1}^{k=n} y_k$$

Propriété :

$$\sum_{k=1}^{k=n} (x_k + y_k)^2 = \sum_{k=1}^{k=n} (x_k{}^2 + y_k{}^2 + 2x_k y_k) = \sum_{k=1}^{k=n} x_k{}^2 + \sum_{k=1}^{k=n} y_k{}^2 + 2\sum_{k=1}^{k=n} x_k y_k$$

3 Variables statistiques quantitatives et qualitatives

Dans le tableau 1.2, nous avons par exemple, le caractère « nombre d'actifs » qui définit la variable statistique X et correspond à un caractère quantitatif ; le caractère « type d'occupation » définit la variable statistique Y et correspond à un caractère qualitatif.

Les individus d'une population sont observés sous l'angle d'un caractère. Une variable statistique permet d'associer à chaque individu de la population une modalité de ce caractère. Une variable peut être de nature quantitative, on parle de variable quantitative ; ou de nature qualitative, on parle de variable qualitative.

Une variable statistique est dite quantitative si ses modalités sont quantifiables. Elle est dite qualitative si ses modalités sont non quantifiables. Les variables quantitatives peuvent être discrètes ou continues : lorsque les modalités numériques d'une variable prennent des valeurs dans l'ensemble N des entiers naturels (respectivement des valeurs dans l'ensemble R des réels) la variable est dite discrète (respectivement la variable est dite continue). Les variables qualitatives peuvent être ordinales ou nominales : lorsque les modalités d'une variable possèdent une relation d'ordre (respectivement ne possèdent pas une relation d'ordre) la variable est dite ordinale (respectivement la variable est dite nominale).

Le regroupement en classes est effectué sur des variables statistiques quantitatives chaque fois que le nombre de valeurs numériques prises par la variable est trop élevé ; cela permet un calcul aisé et rapide. On dit que la variable statistique est groupée en classes.

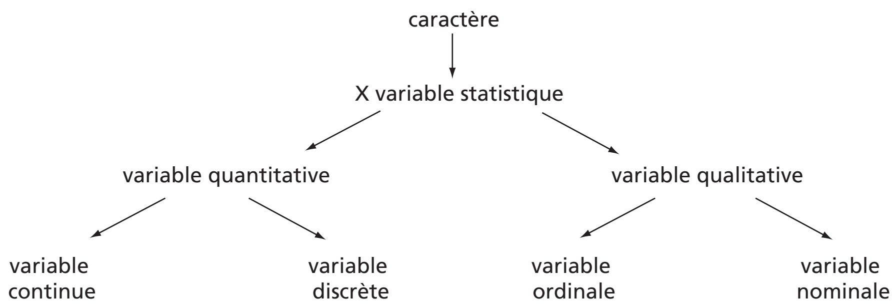


Figure 1.1 : **Caractère et variable statistique**

4 Séries statistiques. Distributions statistiques. Effectifs. Effectif total

Dans le tableau 1.2, le caractère « nombre d'actifs » définit la variable statistique X. La liste des 20 individus (taille de l'échantillon) qui ont répondu à ce caractère donne la série statistique du caractère. Dans le tableau 1.1, le caractère « âge » permet de définir une variable statistique Y ordonnée et groupée en classes, on parle alors de distribution statistique de la variable Y dont le caractère est « âge ».

L'effectif n_i d'une modalité i correspond au nombre d'individus qui possèdent cette modalité. L'effectif d'une modalité est également appelé fréquence absolue. L'effectif total N est égal à la somme des effectifs n_i.

$$N = n_1 + n_2 + \ldots + n_i + \ldots + n_r = \sum_{i=1}^{i=r} n_i$$

Ainsi, le couple (x_i, n_i) représente la distribution de la variable X où n_i est l'effectif et x_i la modalité du caractère (ou réponse) associé à la variable statistique X. L'indice r représente le nombre de réponses (modalités) que l'on peut faire sur le caractère étudié.

La liste des N observations faites pour un caractère d'une population P de taille N est appelée série statistique. Une série statistique ordonnée porte le nom de distribution statistique. Une distribution statistique est l'ensemble des couples (x_i, n_i).

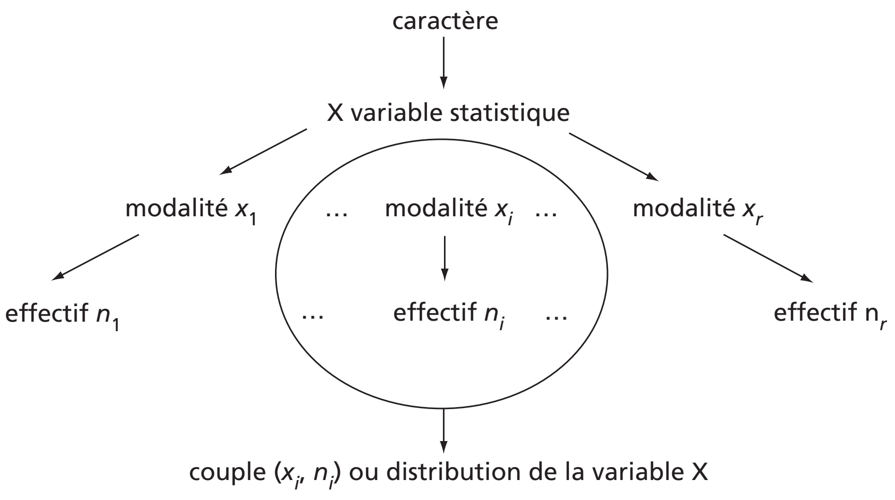


Figure 1.2 : **Variable. Modalités. Effectifs. Distribution statistique**

COURS

5 Séries chronologiques

Une variable statistique quantitative continue ou discrète X est dite temporelle lorsque les modalités x_i font référence au temps t_i. Les prises de mesure x_i sur la variable se font à intervalles réguliers et bien sûr une seule fois au cours du temps t_i.

6 Effectifs cumulés. Fréquences. Fréquences cumulées

6.1 Effectifs cumulés

La distribution en effectifs cumulés (x_i, N_i), présente la somme des effectifs ayant au plus une modalité donnée x_i :

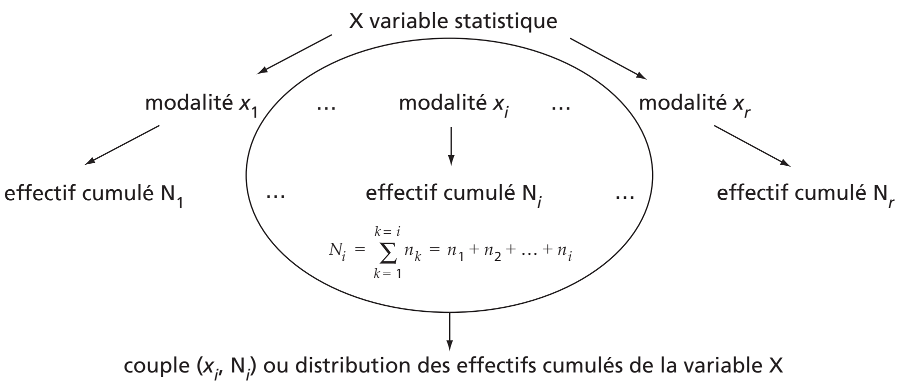


Figure 1.3 : **Distribution en effectifs cumulés**

6.2 Fréquences

La fréquence f_i associée à une modalité i est égale au rapport de l'effectif n_i par l'effectif total N : $f_i = \dfrac{n_i}{N}$

La fréquence est également appelée fréquence relative. Si l'on multiplie par 100 les fréquences, nous obtenons des fréquences en pourcentage notées f_i %.

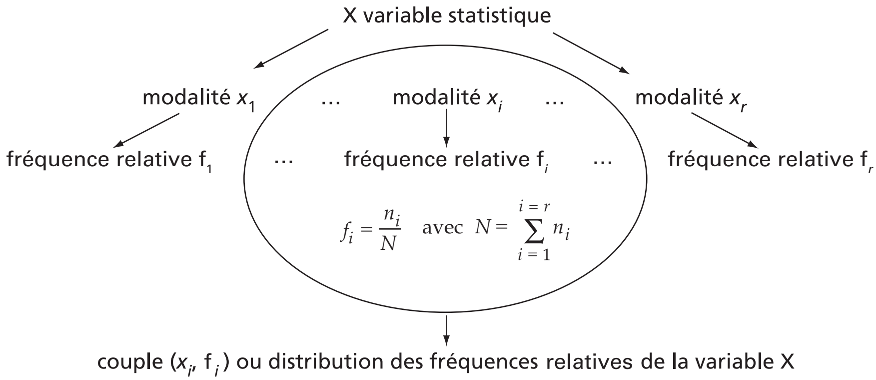


Figure 1.4 : **Distribution en fréquences**

La somme des fréquences est égale à 1 :

$$\sum_{i=1}^{i=r} f_i = \sum_{i=1}^{i=r} \frac{n_i}{N} = \frac{1}{N}\sum_{i=1}^{i=r} n_i = \frac{1}{N}N = 1$$

6.3 Fréquences cumulées

Les distributions en fréquences cumulées (x_i, F_i) présentent la somme des fréquences ayant au plus une modalité donnée x_i.

Si l'on multiplie par 100 les fréquences cumulées, nous obtenons des fréquences cumulées en pourcentage notées F_i %.

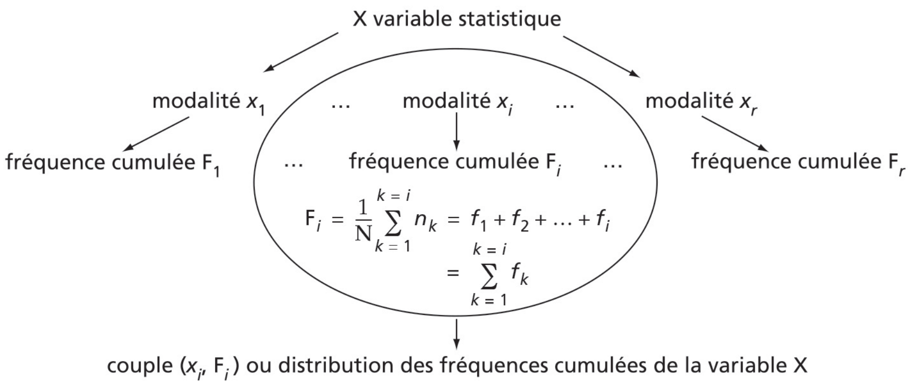


Figure 1.5 : **Distribution en fréquences cumulées**

COURS

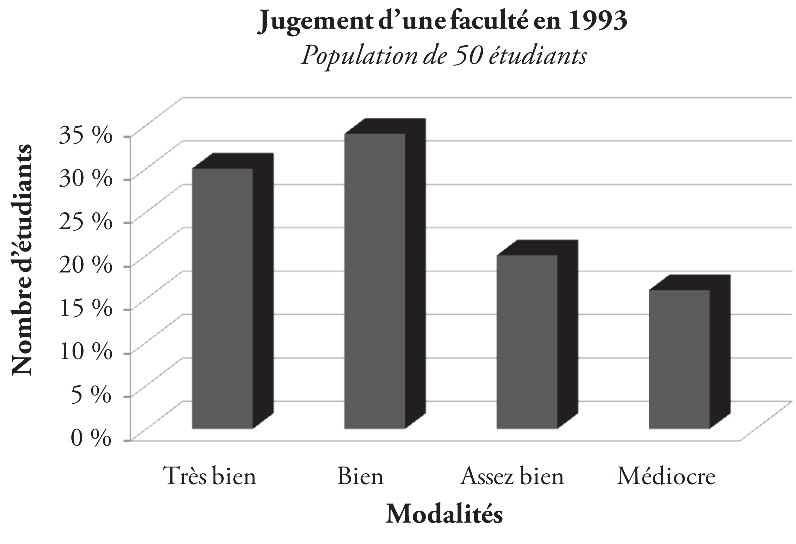
7 Principaux graphiques des variables statistiques

Pour résumer l'information contenue dans une distribution statistique (quantitative ou qualitative), on utilise en premier lieu des résumés graphiques. Chaque type de variable possède une ou plusieurs représentations graphiques.

7.1 Représentations graphiques des variables qualitatives

▸ Le diagramme en « tuyaux d'orgue »

Dans ce type de représentation, les rectangles (tuyaux) ont pour base une modalité et comme hauteur l'effectif (ou la fréquence). La base de chacun des rectangles (base sur l'axe des abscisses) ne possède aucune signification numérique puisque la variable est qualitative.

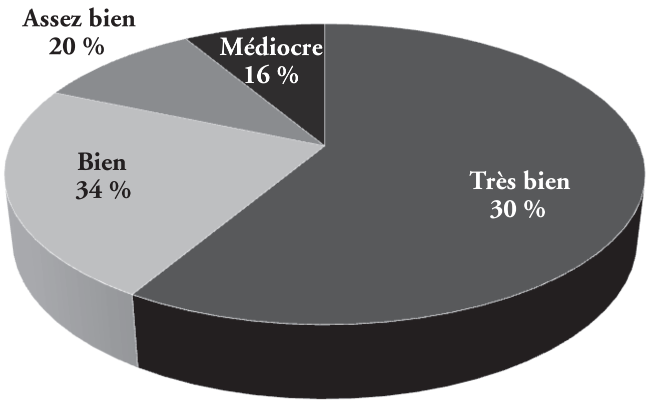


▸ Le diagramme en secteurs

Dans ce type de représentation, nous utilisons un disque (plus communément appelé camembert). Chacune des modalités est représentée par un secteur qui est proportionnel à l'effectif (ou à la fréquence).

Jugement d'une faculté en 1993

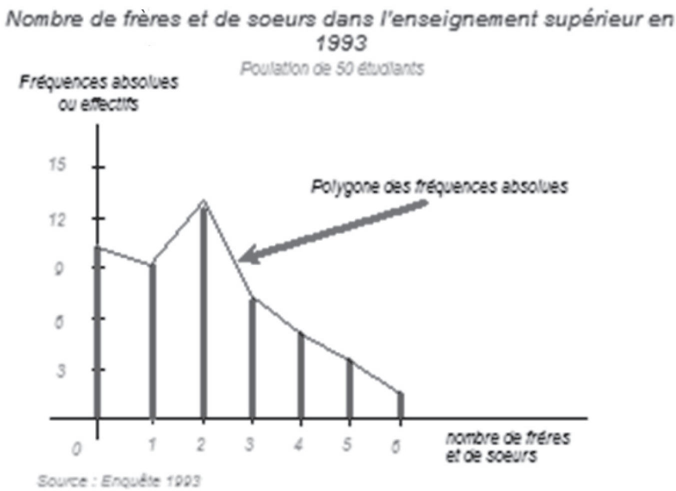
Population de 50 étudiants



7.2 Représentations graphiques des variables quantitatives

▸ Le diagramme bâtons

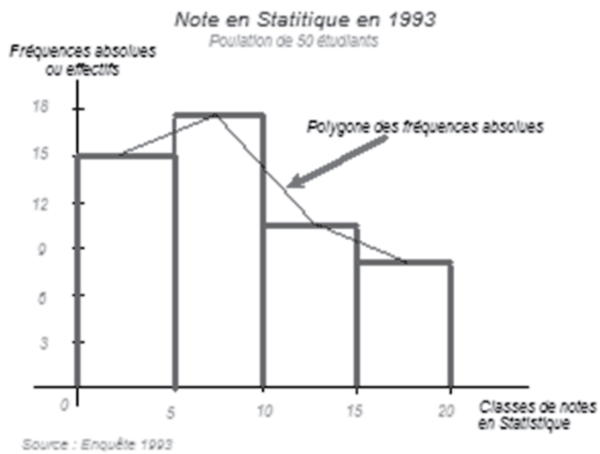
Le diagramme bâtons est la représentation d'une distribution statistique discrète (x_i, n_i) ou (x_i, f_i) ou $(x_i, f_i\%)$, où chaque bâton a une hauteur proportionnelle à l'effectif n_i ou à la fréquence f_i (ou $f_i\%$) de la modalité x_i.



▸ L'histogramme

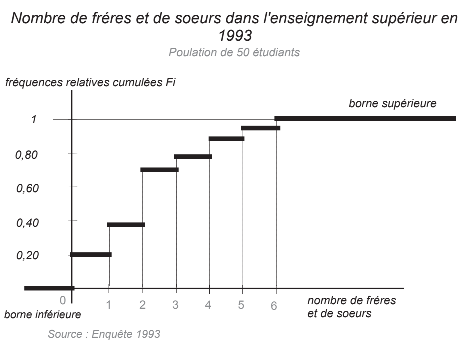
L'histogramme est la représentation graphique d'une distribution statistique (x_i, n_i) ou (x_i, f_i) ou $(x_i, f_i\%)$ groupée en classes, où chaque classe est représentée par un rectangle de base proportionnelle à l'amplitude et de surface proportionnelle à l'effectif ou à la fréquence.

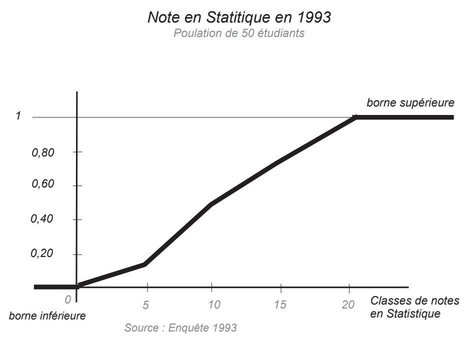
Les distributions statistiques de variables quantitatives groupées en classes peuvent avoir des amplitudes différentes. Dans ce cas, c'est à partir de distributions statistiques corrigées que l'on construit un histogramme.



▸ Le diagramme cumulatif croissant

Le diagramme cumulatif croissant est la représentation graphique de la distribution des fréquences cumulées (x_i, F_i) (ou des effectifs cumulés (x_i, N_i) ou des fréquences cumulées en pourcentage (x_i, F_i%)).





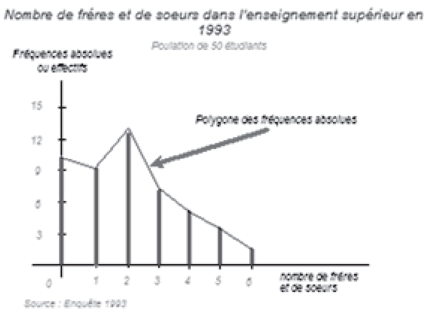
Le diagramme cumulatif croissant d'une variable quantitative discrète est une fonction dite en « escalier ». Cette fonction est par hypothèse continue à gauche.

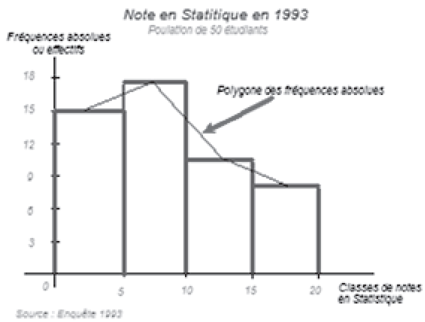
La courbe du diagramme cumulatif d'une variable groupée en classes passe par les limites supérieures des classes.

▸ Le polygone des fréquences

Le polygone des fréquences d'une variable discrète est la courbe qui joint les sommets des bâtons d'un diagramme bâtons.

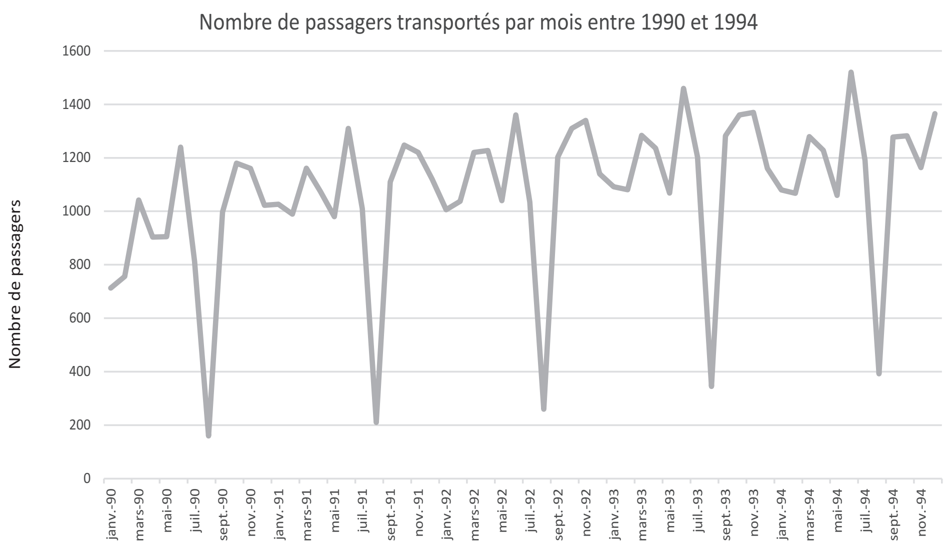
Le polygone des fréquences d'une variable continue groupée en classes est la courbe qui joint les milieux des côtés supérieurs des rectangles d'un histogramme.





7.3 Séries chronologiques

Pour représenter une série chronologique, nous devrions utiliser un diagramme bâtons, mais la lisibilité est moins bonne que lorsqu'on utilise une courbe qui joint les points dont les coordonnées sont, en abscisse, le temps, et en ordonnée, la valeur de la variable.



Q.C.M.

	Vrai	Faux
1. La statistique descriptive décrit et analyse.	○	○
2. Une variable statistique discrète ne peut pas être groupée en classes.	○	○
3. Une variable statistique continue possède des valeurs dans l'ensemble des naturels.	○	○
4. Un histogramme est la représentation graphique d'une variable statistique continue groupée en classes.	○	○
5. L'aire d'un rectangle composant un histogramme est proportionnelle à l'effectif de la classe.	○	○
6. On ne peut pas calculer des fréquences cumulées croissantes sur une variable statistique qualitative ordinale.	○	○
7. La représentation d'une variable statistique discrète est le diagramme en bâtons.	○	○
8. Le polygone des fréquences ou des effectifs permet de représenter une variable qualitative nominale.	○	○
9. L'opérateur sigma sert à contracter l'opération multiplication.	○	○
10. Une variable statistique est toujours associée à un caractère.	○	○

Réflexion

11. Calculer les expressions suivantes :

$$\sum_{i=1}^{i=n}\left(x_i - a\right)^2 \qquad \sum_{i=1}^{i=n}\left(x_i - y_i\right)^2$$

12. En préparation au calcul de la covariance, que nous définirons dans les TD suivants, calculer la quantité :

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}\left(x_i - a\right)\cdot\left(y_j - b\right)$$

Donner l'expression quand a = b et i = j.

ENTRAÎNEMENT

13. Construction des tableaux statistiques

Le tableau 1.2 est un extrait de la base de données sur le comportement des consommateurs de la région de Nantes, enquête réalisée en juin 1999. Après avoir pris connaissance de ce tableau individus-caractères, répondre à la question suivante : pour chacun des caractères « âge », « nombre de voitures », « type d'occupation » du logement**,** définir sa nature et construire sa distribution statistique sous forme de tableau.

Analyse de l'énoncé et conseils. Cet exercice vous permettra de prendre conscience que les tableaux de la statistique sont issus d'hypothèses et de définitions. Compter et regrouper en fonction des définitions sur les variables statistiques sont les buts recherchés.

14. Variable qualitative nominale

Une enquête sur 842 commerces de la région Languedoc-Roussillon en collaboration avec la CCI Languedoc-Roussillon a permis de présenter une étude sur les « Soldes ETE 2016 ». Le tableau ci-dessous donne la réparation des enquêtes sur les neuf pôles commerces du Languedoc-Roussillon :

Tableau 1.3 : **Répartition des enquêtes commerces sur les soldes été 2016**

Pôles de l'enquête	Nombre d'enquête
Alès	64
Béziers	48
Carcassonne	78
Mende	17

EXERCICES

Pôles de l'enquête	Nombre d'enquête
Montpellier	235
Narbonne	56
Nîmes	136
Perpignan	129
Sète	79
Total général	842

1. Quelle est la nature de ce caractère ? Définir la variable statistique associée à ce caractère.

2. Calculer l'effectif total et les fréquences de cette distribution.

3. Donner plusieurs représentations graphiques de cette variable.

Analyse de l'énoncé et conseils. Le caractère est qualitatif nominal, son traitement statistique est « limité » à l'étude des fréquences. Il est à noter qu'un tel caractère possède plusieurs représentations graphiques.

15. Variable qualitative ordinale

La distribution des 842 commerces du Languedoc-Roussillon suivant la variable « Évolution du chiffre d'affaires » de 2016 par rapport à 2015, est la suivante :

Tableau 1.4 : **Répartition des enquêtes par nomenclature de commerces**

Évolution du chiffre d'affaires	Nombre de commerces
Évolution négative	415
Stabilité	156
Évolution positive	271
Total général	842

1. Quelle est la nature de ce caractère ? Définir la variable statistique associée à ce caractère.

2. Calculer l'effectif total et les fréquences de cette distribution.

3. Donner plusieurs représentations graphiques de cette variable.

Analyse de l'énoncé et conseils. Les variables statistiques ordonnées ont des représentations graphiques différentes. Les calculs sur ce type de variable ne portent que sur les effectifs ou les fréquences (simples et cumulées) de chacune des modalités.

16. L'opérateur ($\sum$) et l'opérateur ($\prod$)

On vous propose le tableau suivant :

Numéro	X	Y	Constante
1	12	84	2
2	3	73	2
3	44	74	2
4	10	59	2
5	57	24	2
6	76	95	2
7	77	82	2
8	98	88	2

On vous demande de calculer les quantités suivantes :

1. La somme des valeurs de X.

2. La sommes des valeurs de Y.

3. La somme de la constante.

4. La somme des carrés de X, puis la somme des carrés de Y.

5. Le carré de la somme de X plus Y.

6. La somme du produit des carrés des deux variables X et Y.

7. La somme de la somme des variables X, Y et de la constante.

8. La somme du produit des variables X, de Y et de la constante.

9. Comparer les résultats obtenus lors des calculs des questions.

10. La somme du produit des variables X et Y.

11. La somme du produit des carrés de X et Y.

Toutes ces quantités seront calculées sur tableau et en utilisant les opérateurs somme et produit.

Analyse de l'énoncé et conseils. Cet exercice est important dans la mesure où il vous permettra de comprendre l'utilisation des opérateurs mathématiques l'on doit connaître pour comprendre les formules et les démonstrations faites en statistique descriptive.

17. Variable quantitative discrète

La répartition de 2205 ménages dans une commune en fonction du nombre de voitures dans le foyer est donnée dans le tableau ci-dessous.

Tableau 1.5 : **Distribution du nombre de voitures par ménage**

Nombre de voitures dans le ménage	Nombre de ménages
0	209
1	1072
2	824
3	84
4	14
5	2
Total général	2205

1. Donner la nature du caractère « taille du ménage » et définir la variable statistique associée.

2. Calculer les effectifs, les fréquences, les fréquences cumulées.

3. Tracer le graphique correspondant à la variable statistique étudiée. Représenter le polygone des fréquences.

Analyse de l'énoncé et conseils. Cet exercice vous permettra de tracer le graphique qui correspond à la variable statistique et de voir qu'il existe qu'une seule représentation possible pour ce style de variable. Une variable statistique discrète est représentée obligatoirement par un « diagramme en bâtons ». Le fait de connaître la dénomination de la variable, entraîne obligatoirement le type de graphique.

18. Variable quantitative continue

Dans une commune il est demandé à 1556 ménages de donner leur revenu, et pour cela on présente un formulaire sous forme de classes prédéfinies, voir le tableau 1.6.

Tableau 1.6 : **Répartition du nombre de ménages dans 7 classes de revenu**

Classes	Borne inférieure de revenu	Borne supérieure de revenu	Nombre de ménages
Classe 1	0	5823	281
Classe 2	5823	11661	661
Classe 3	11661	17484	402
Classe 4	17484	23307	148
Classe 5	23307	29130	44
Classe 6	29130	34953	34
			1570

1. Calculer les fréquences et donner la représentation graphique. Tracer le polygone des fréquences.

2. Déterminer les fréquences cumulées croissantes et décroissantes et les représenter graphiquement.

Analyse de l'énoncé et conseils. Les variables statistiques quantitatives continues peuvent prendre toutes les valeurs sur l'ensemble des réels. La principale difficulté de cet exercice réside dans la construction d'un graphique appelé histogramme, avec des amplitudes de classes différentes ou égales.

SOLUTIONS

1. Faux
2. Faux
3. Vrai
4. Vrai
5. Vrai
6. Faux
7. Vrai
8. Faux
9. Faux
10. Vrai

11. Calcul des expressions suivantes :

Première expression : $\sum_{i=1}^{i=n} (x_i - a)^2$

Développons l'expression :

$$\sum_{i=1}^{i=n}(x_i - a)^2 = \sum_{i=1}^{i=n}\left(x_i^2 + a^2 - 2\cdot a\cdot x_i\right)$$

Distribuons la somme à l'intérieure des parenthèses :

$$\sum_{i=1}^{i=n}(x_i - a)^2 = \sum_{i=1}^{i=n}\left(x_i^2 + a^2 - 2\cdot a\cdot x_i\right)$$

$$\sum_{i=1}^{i=n}(x_i - a)^2 = \sum_{i=1}^{i=n}x_i^2 + \sum_{i=1}^{i=n}a^2 - \sum_{i=1}^{i=n}2\cdot a\cdot x_i$$

$$\sum_{i=1}^{i=n}(x_i - a)^2 = \sum_{i=1}^{i=n}x_i^2 + a^2\sum_{i=1}^{i=n}1 - 2\cdot a\sum_{i=1}^{i=n}x_i$$

$$\sum_{i=1}^{i=n}(x_i - a)^2 = \sum_{i=1}^{i=n}x_i^2 + a^2 n - 2\cdot a\sum_{i=1}^{i=n}x_i$$

Deuxième expression : $\sum_{i=1}^{i=n}(x_i - y_i)^2$

$$\sum_{i=1}^{i=n}(x_i - y_i)^2 = \sum_{i=1}^{i=n}\left(x_i^2 + y_i^2 - 2\cdot x_i\cdot y_i\right)$$

$$\sum_{i=1}^{i=n}(x_i - y_i)^2 = \sum_{i=1}^{i=n}x_i^2 + \sum_{i=1}^{i=n}y_i^2 - 2\sum_{i=1}^{i=n}x_i\cdot y_i$$

12. En préparation au calcul de la covariance, que nous définirons dans les TD suivants, calculer la quantité :

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}(x_i - a)\cdot(y_j - b)$$

Multiplions les deux polynômes :

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}(x_i - a)\cdot(y_j - b) = \sum_{j=1}^{j=r}\sum_{i=1}^{i=n}\left[x_i\cdot y_j - x_i\cdot b - a\cdot y_j + a\cdot b\right]$$

Distribuons le terme double somme :

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_j-b)=\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}x_i\cdot y_j-\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}x_i\cdot b-\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}a\cdot y_j+\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}a\cdot b$$

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_j-b)=\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}x_i\cdot y_j-b\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}x_i\cdot-a\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}\cdot y_j+a\cdot b\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}1$$

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_j-b)=\sum_{j=1}^{j=r}y_j\sum_{i=1}^{i=n}x_i-b\sum_{i=1}^{i=n}x_i\sum_{j=1}^{j=r}1-a\sum_{j=1}^{j=r}y_j\sum_{i=1}^{i=n}1+a\cdot b\sum_{j=1}^{j=r}n$$

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_j-b)=\sum_{j=1}^{j=r}y_j\sum_{i=1}^{i=n}x_i-b\sum_{i=1}^{i=n}x_i r-a\sum_{j=1}^{j=r}y_j n+a\cdot b\cdot n\sum_{j=1}^{j=r}1$$

$$\sum_{j=1}^{j=r}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_j-b)=\sum_{j=1}^{j=r}y_j\sum_{i=1}^{i=n}x_i-b\cdot r\sum_{i=1}^{i=n}x_i-a\cdot n\sum_{j=1}^{j=r}y_j+a\cdot b\cdot n\cdot r$$

Quand a = b et i = j :

$$\sum_{i=1}^{i=n}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_i-a)=\sum_{i=1}^{i=n}y_i\sum_{i=1}^{i=n}x_i-a\cdot n\sum_{i=1}^{i=n}x_i-a\cdot n\sum_{i=1}^{i=n}y_i+a^2\cdot n^2$$

$$\sum_{i=1}^{i=n}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_i-a)=\sum_{i=1}^{i=n}y_i\sum_{i=1}^{i=n}x_i-a\cdot n\left(\sum_{i=1}^{i=n}x_i+\sum_{i=1}^{i=n}y_i\right)+(a\cdot n)^2$$

$$\sum_{i=1}^{i=n}\sum_{i=1}^{i=n}(x_i-a)\cdot(y_i-a)=\sum_{i=1}^{i=n}y_j\sum_{i=1}^{i=n}x_i-a\cdot n\sum_{i=1}^{i=n}(x_i+y_i)+(a\cdot n)^2$$

13. Construction des tableaux statistiques

(Voir également le TD Application Excel 8, Construction de tableaux, p. 27.)

1) Le caractère « âge » doit être associé à une variable statistique quantitative continue. Afin de simplifier la construction du tableau statistique sur l'âge des 20 individus, nous regrouperons en classes la variable, en prenant en compte les différents éléments de construction : le nombre d'individus de la population ou de l'échantillon, ainsi que le nombre de classes que l'on désire créer (le nombre de classes dépend de plusieurs paramètres, mais nous supposerons ici qu'il est fixé de façon arbitraire pour simplifier la démarche).

Le nombre de classes est fixé à 4. On dit alors que les 20 individus de l'échantillon sont regroupés en 4 classes.

Les bornes des classes : ce sont les limites inférieures et supérieures de classe ; seule une des deux bornes est comprise (problème de continuité) :
première classe : de 20 ans à moins de 40 ans
– écriture sous forme d'intervalle [20 ; 40 [
– borne inférieure : 20 ans borne comprise
– borne supérieure : 40 ans borne non comprise
seconde classe : de 40 ans à moins de 60 ans
– écriture sous forme d'intervalle [40 ; 60 [
– borne inférieure : 40 ans borne comprise
– borne supérieure : 60 ans borne non comprise
troisième classe : de 60 ans à moins de 75 ans
– écriture sous forme d'intervalle [60 ; 75 [
– borne inférieure : 60 ans borne comprise
– borne supérieure : 75 ans borne non comprise
quatrième classe : 75 ans et plus
– écriture sous forme d'intervalle [75 ; plus de 75 ans [
– borne inférieure : 75 ans borne comprise
– borne supérieure ouverte. Dans ce cas, il est nécessaire de borner la classe lorsqu'on désire effectuer des calculs.

Les amplitudes des classes : ce sont les longueurs de chaque classe, ou les différences entre les bornes supérieures et inférieures :
première classe : [20 ; 40 [, l'amplitude est de $a_1 = 40 - 20 = 20$
seconde classe : [40 ; 60 [, l'amplitude est de $a_2 = 60 - 40 = 20$
troisième classe : [60 ; 75 [, l'amplitude est de $a_3 = 75 - 60 = 15$
quatrième classe : [75 ; plus de 75 ans [, pas d'amplitude. Elle est dans ce dernier cas infinie. La borne supérieure de classe est ouverte : il est nécessaire de fixer une valeur ; prenons par exemple 100 ans. Dans ce cas, l'amplitude de la classe devient : $a_1 = 100 - 75 = 25$.

Le regroupement en classes que l'on vient de réaliser est un regroupement à classes d'amplitudes inégales.

Les centres des classes : ce sont les milieux de chaque classe, ils résument toutes les valeurs contenues dans la classe. On fait l'hypothèse que les individus d'une même classe ont même valeur, la valeur centrale (on dit que la distribution est uniforme à l'intérieur de chacune des classes).
première classe : [20 ; 40 [, le centre de la classe est de $(20 + 40)/2 = 30$
seconde classe : [40 ; 60 [, le centre de la classe est de $(40 + 60)/2 = 50$
troisième classe : [60 ; 75 [, le centre de la classe est de $(60 + 75)/2 = 67{,}5$
quatrième classe : [75 ; 100 [, le centre de la classe est de $(75 + 100)/2 = 87{,}5$