IDUP/Master 1/Analyse statistique

# 1. LECTURE DE TABLEAUX STATISTIQUES

*Formalisation et Définitions de base*

## Population statistique

Une *population* statistique, c'est l'ensemble des éléments sur lesquels porte l'étude. Il s'agit du « périmètre » de l'étude.

Ces éléments sont les *individus* statistiques (ou unités statistiques).

En démographie, les individus sont des personnes, la population étant un ensemble de personnes qui ont au moins une caractéristique commune.

En général, on note N l'effectif total de la population.

sept.-20 JF Léger - IDUP 2

Exemples (cf. Tableaux distribués en cours) :

- la population de la France métropolitaine : c'est l'ensemble des personnes résidant en France. Elle compte en 2006 61 795 006 individus (tableau 6).

- la population de Lyon, ce sont toutes les personnes qui ont pour caractéristique commune de résider dans cette commune. Du point de vue statistique, chaque habitant de la commune est un individu statistique. En 2006, la ville de Lyon en compte 472 330 (tableau 3).

On compte aussi des populations non humaines, comme le parc de logements.

En France métropolitaine, la population des logements est de 26 696 843 en 2006 (tableau 5). A Lyon, on compte 236 819 logements en 2006 (tableau 2).

**« Périmètre » de l'étude** :

L'INSEE définit plusieurs types de population : la population légale, la population municipale, la population des ménages, etc. Selon la définition retenue, la taille de la population d'une ville varie (cf. INSEE Première n° 1217, janvier 2009). Quand on fait des comparaisons (dans le temps, ou entre villes à un même moment), il convient donc de bien préciser le type de population retenue afin de proposer des comparaisons à « périmètre » constant.

## Echantillon

Un *échantillon* est un sous ensemble de la population.

Sa taille est inférieure à celle de la population dont il est issu (appelée population mère).

## Variable statistique

Une *variable* statistique (ou caractéristique statistique) est un critère qui va permettre de classer les individus, de les répartir, de les distinguer les uns des autres.

Par rapport à ce critère, à cette variable, les individus vont être placés dans des groupes, chaque groupe correspondant à une des valeurs ou une des formes que peut prendre le critère considéré. Celles-ci sont appelées des *modalités*.

Notation : en général, les variables sont notées X, Y, etc.

sept.-20 JF Léger - IDUP 4

**Exemple de variables** :

(1) Pour une population humaine : le sexe, l'âge, le lieu de résidence cinq ans auparavant, la CSP, etc.

(2) Pour une population de logements : la catégorie, le type, le nombre de pièces, la date d'achèvement, etc.

**Exemple de modalités** :

- Pour la variable sexe, il existe deux modalités : Homme, Femme

- Pour la variable âge : les 0-4 ans, les 5-9 ans, etc.

## Caractéristiques des modalités

Les *modalités* répondent à deux conditions absolument nécessaires pour classer correctement les individus selon la variable retenue :

(1) Elles sont exhaustives, à savoir qu'elles doivent recouvrir toutes les possibilités envisageables. Une modalité doit pouvoir être affectée à chaque individu.

(2) Elles sont exclusives les unes des autres : deux modalités ne peuvent être attribuées au même individu.

Notation : les modalités sont notées $x_i$, $y_i$, etc.

sept.-20 JF Léger - IDUP 5

Exemples :

- par rapport à la variable sexe, un individu est classé parmi les hommes OU les femmes ;

- par rapport à la variable âge, un individu appartient à UNE ET UNE seule classe d'âges.

## Types de variables

Il existe deux types de variables :

(1) Les *variables qualitatives* : une variable est qualitative si ses modalités ne sont pas mesurables.

Exemples : la PCS, le sexe, le lieu de résidence cinq ans auparavant.

(2) Les *variables quantitatives* :

Une variable est quantitative quand ses modalités sont mesurables.

Exemples : l'âge, l'année d'achèvement des résidences principales d'un parc immobilier, le nombre de pièces par ménage, etc.

## Variables qualitatives

On peut distinguer les variables qualitatives *nominales* des variables qualitatives *ordinales*.

(1) Variable qualitative *nominale* : une variable qualitative est nominale lorsque les modalités ne sont pas naturellement ordonnées.

(2) Variable qualitative *ordinale* : au contraire une variable qualitative est ordinale lorsque l'ensemble de ses catégories (les modalités) peuvent être hiérarchisées, c'est-à-dire classées les unes par rapport aux autres.

Représentations graphiques associées : le diagramme en tuyaux d'orgue (ou diagramme en barres), le diagramme circulaire ou semi-circulaire.

sept.-20 JF Léger - IDUP 7

**Exemple de variables qualitatives nominales** :

- La catégorie socio professionnelle ou le sexe. Les différentes modalités de chacune de ces variables ne sont pas hiérarchisées les unes par rapport aux autres.

**Exemple de variables qualitatives ordinales** :

Le diplôme est une variable qualitative (elle n'est pas mesurable) ordinale (on peut hiérarchiser les catégories – modalités). On peut par exemple classer par ordre croissant ou décroissant les niveaux de diplôme en fonction de la hiérarchie scolaire.

## Variables quantitatives

Il existe deux types de variables quantitatives :

(1) Les variables quantitatives *discrètes* : une variable quantitative est discrète lorsqu'elle prend un nombre fini de valeurs (de modalités).

Représentation graphique associée : le diagramme en bâtons.

(2) Les variables quantitatives *continues* : Une variable quantitative est continue lorsqu'elle comporte un nombre potentiellement infini de valeurs (ou modalités). Dit autrement, elle peut prendre un nombre infini de valeurs sur un intervalle donné.

Représentation graphique associée : l'histogramme.

sept.-20 JF Léger - IDUP 8

**Exemple de variable quantitative discrète** :

- Le nombre de pièces pour la population des résidences principales. Un logement a 1, 2, 3, etc. pièces. On peut certes calculer des caractéristiques de tendance centrale comme le nombre moyen de pièces par logement (qui peut comporter des décimales), mais les valeurs qui conduisent à ce résultat sont des nombres entiers.

- C'est aussi le cas par exemple quand on distribue une population de femmes ou de ménages en fonction du nombre d'enfants : chaque femme a un nombre entier d'enfants (en tout cas, on l'espère !). La moyenne du nombre d'enfants par femme peut en revanche comporter des décimales.

**Exemple de variable quantitative continue** :

C'est par exemple le cas de l'âge. Certes la vie humaine est limitée, mais entre 0 et 1 an, on peut exprimer les âges en mois, en jours, en heures, minutes, etc. On peut donc classer les individus selon des échelles d'âge de précisions diverses. Selon le degré de précision, l'âge peut prendre un nombre de modalités quasiment infini.

Pour étudier une variable quantitative continue, on définit le plus souvent des classes ou des intervalles afin de travailler sur des tableaux réduits ou/et de commencer à synthétiser l'information. On rend discrète, en quelque sorte, la variable continue : on la discrétise.

C'est par exemple ce que l'on fait pour la variable âge quand on présente la distribution d'une population selon l'âge. On peut certes classer les individus selon l'âge détaillé (année d'âge) mais aussi par groupe d'âges quinquennaux (on regroupe cinq années d'âge), décennaux, etc.

## Formalisation d'un tableau « simple »

Distribution de la population lyonnaise selon l'âge

| Age semi-détaillé | Ensemble |
|---|---|
| Moins de 3 ans | 17 637 |
| 3 à 5 ans | 15 051 |
| 6 à 10 ans | 21 545 |
| 11 à 17 ans | 30 719 |
| 18 à 24 ans | 74 908 |
| 25 à 39 ans | 119 428 |
| 40 à 54 ans | 79 363 |
| 55 à 64 ans | 44 732 |
| 65 à 79 ans | 44 914 |
| 80 ans ou plus | 24 035 |
| Ensemble | 472 332 |

Distribution d'une population selon la variable X

| X | $n_i$ |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| $x_3$ | $n_3$ |
| $x_4$ | $n_4$ |
| $x_5$ | $n_5$ |
| $x_6$ | $n_6$ |
| $x_7$ | $n_7$ |
| $x_8$ | $n_8$ |
| $x_9$ | $n_9$ |
| $x_{10}$ | $n_{10}$ |
| Total | N |

sept.-20 JF Léger - IDUP 9

D'une manière générale, on note X la variable qui va permettre de distribuer (effectifs) ou de répartir (proportions) la population étudiée.

Chaque modalité de la variable X est notée $x_i$, i variant de 1 à *n*, *n* étant le nombre de modalités. Dans le cas présent, il y a 10 modalités, i varie donc de 1 à 10.

Les effectifs correspondant à chaque modalité $x_i$ sont notées $n_i$. La somme des effectifs correspondant aux différentes modalités correspond à la taille de la population étudiée, notée N.

Dans l'exemple ci-dessus, X correspond à l'âge semi détaillé. La modalité $x_1$ correspond au groupe d'âges 0-2 ans révolus (Moins de 3 ans); la modalité $x_2$ correspond au groupe d'âges 3-5 ans, etc.

L'effectif de personnes âgées de 11 à 17 ans correspond à $n_4$ dans le tableau général.

## Fréquences

On distingue deux types de fréquence dans un tableau « simple » :

(1) La fréquence *absolue* : il s'agit de l'effectif contenu dans une cellule du tableau. Il s'agit du nombre de personnes ($n_i$) qui ont pour point commun une modalité particulière ($x_i$) de la variable X.

La somme des effectifs correspond à la taille de la population ou effectif total N.

$$N = \sum_{i=\alpha}^{\omega} n_i$$

(2) La fréquence (ou fréquence *relative*) : la fréquence désigne la proportion ($f_i$) d'individus, parmi l'ensemble des individus composant la population qui fait l'objet de l'étude (N) qui sont affectés à la modalité $x_i$ de la variable X.

La somme des fréquences est égal à 1 ou 100 %.

$$\sum_{i=\alpha}^{\omega} f_i = \sum_{i=\alpha}^{\omega} \frac{n_i}{N} = \frac{1}{N} \sum_{i=\alpha}^{\omega} n_i = 100\%$$

## Formalisation du tableau de contingence

Distribution de la population lyonnaise selon l'âge et le sexe

| Age semi-détaillé | Sexe | | Ensemble |
|---|---|---|---|
| | Hommes | Femmes | |
| Moins de 3 ans | 9 146 | 8 491 | 17 637 |
| 3 à 5 ans | 7 607 | 7 444 | 15 051 |
| 6 à 10 ans | 10 745 | 10 800 | 21 545 |
| 11 à 17 ans | 15 580 | 15 139 | 30 719 |
| 18 à 24 ans | 32 650 | 42 258 | 74 908 |
| 25 à 39 ans | 59 979 | 59 449 | 119 428 |
| 40 à 54 ans | 37 682 | 41 681 | 79 363 |
| 55 à 64 ans | 21 038 | 23 694 | 44 732 |
| 65 à 79 ans | 18 047 | 26 867 | 44 914 |
| 80 ans ou plus | 7 250 | 16 785 | 24 035 |
| Ensemble | 219 724 | 252 608 | 472 332 |

Distribution d'une population selon les variables X et Y

| X | Y | | Ensemble |
|---|---|---|---|
| | $y_1$ | $y_2$ | |
| $x_1$ | $n_{1,1}$ | $n_{1,2}$ | $n_{1.}$ |
| $x_2$ | $n_{2,1}$ | $n_{2,2}$ | $n_{2.}$ |
| $x_3$ | $n_{3,1}$ | $n_{3,2}$ | $n_{3.}$ |
| $x_4$ | $n_{4,1}$ | $n_{4,2}$ | $n_{4.}$ |
| $x_5$ | $n_{5,1}$ | $n_{5,2}$ | $n_{5.}$ |
| $x_6$ | $n_{6,1}$ | $n_{6,2}$ | $n_{6.}$ |
| $x_7$ | $n_{7,1}$ | $n_{7,2}$ | $n_{7.}$ |
| $x_8$ | $n_{8,1}$ | $n_{8,2}$ | $n_{8.}$ |
| $x_9$ | $n_{9,1}$ | $n_{9,2}$ | $n_{9.}$ |
| $x_{10}$ | $n_{10,1}$ | $n_{10,2}$ | $n_{10.}$ |
| Total | $n_{.1}$ | $n_{.2}$ | $n_{..} = N$ |

sept.-20 JF Léger - IDUP 11

Les individus de la population étudiée sont distribués (ou répartis) en fonction de deux variables nommés classiquement X et Y.

- La variable X présente *n* modalités notées $x_i$. Ici, *n* vaut 10. Dans le tableau, les modalités sont placées les unes sous les autres dans une même colonne.

- La variable Y présente *p* modalités notées $y_j$. Ici, *p* est égal à 2. Il faut noter que le nombre de modalités de la variable X peut être égal à celui de la variable Y. En conséquence, *n* peut être égal à *p*. Dans le tableau, les modalités de la variable Y sont placées sur une même ligne.

De ce fait, chaque ligne du tableau regroupe les individus ayant en commun une même modalité de la variable X. Et chaque colonne regroupe les individus ayant en commun une même modalité de la variable Y.

Le nombre de personnes caractérisées à la fois par la modalité i de X et j de Y est noté $n_{ij}$. Le premier indice correspond donc toujours à la modalité i de la variable X et le second à la modalité j de la variable Y. Par exemple, $n_{3,1}$ correspond au nombre d'individus caractérisés par la modalité 3 de X et la modalité 1 de Y. Dans le tableau de gauche, cela correspondrait aux Lyonnais âgés de 6 à 10 ans (modalité 3 de la variable X) de sexe masculin (modalité 1 de la variable Y).

La somme des personnes présentant la caractéristique i de la variable X, quelle que soit la valeur prise par la modalité j de Y, est notée $n_{i.}$. Cela signifie que l'on a bloqué la modalité i de X et que l'on a sommé, compte tenu de cette contrainte, tous les individus quelle que soit j de Y. Par exemple, dans le tableau de gauche, $n_{2.}$ correspond à l'ensemble des Lyonnais âgés de 3 à 5 ans, quel que soit leur sexe.

La somme des personnes présentant la caractéristique j de Y, quelle que soit la valeur prise par la modalité i de X, est notée $n_{.j}$. Cela signifie que l'on a bloqué j de Y et que l'on a sommé tous les individus quel que soit i de X. Par exemple, $n_{.2}$ correspond, dans le tableau de gauche, à l'ensemble de la population lyonnaise de sexe féminin, quel que soit l'âge.

$n_{i.}$ et $n_{.j}$ sont appelés « effectifs marginaux » ou « sommes marginales ». Il s'agit des totaux situés dans les marges du tableau de contingence.

La somme des $n_{ij}$ correspond à la somme des $n_{i.}$ Elle-même équivalente à la somme des $n_{.j}$. Elles correspondent chacune à N.

$$\sum_{i=1}^{n} n_{i.} = \sum_{j=1}^{p} n_{.j} = N$$

## Fréquences relatives dans un tableau de contingence

(1) Fréquences relatives élémentaires

$$f_{i,j} = \frac{n_{i,j}}{n..} \; avec \; ici \sum_{i=1}^{10} \sum_{j=1}^{2} f_{i,j} = 1$$

$$ou \sum_{i=1}^{n} \sum_{j=1}^{p} f_{i,j} = 1$$

avec n = nombre de modalités de la variable X

avec p = nombre de modalités de la variable Y

sept.-20 JF Léger - IDUP 12

Les fréquences relatives élémentaires correspondent au poids de chaque catégorie d'individus définie au croisement des deux variables X et Y au sein de la population totale.

Il s'agit donc de la proportion de personnes présentant à la fois les caractéristiques i de X ($x_i$) et j de Y ($y_j$) dans la population totale. Dans l'exemple précédent, $f_{2,1}$ correspondrait à la part des garçons âgés de 3-5 ans au sein de la population totale.

La somme des fréquences relatives élémentaires vaut 1 ou 100 %.

## Fréquences relatives dans un tableau de contingence

(2) Fréquences relatives conditionnelles

*a) en colonne*

$$f_{i/j} = \frac{n_{i,j}}{n._j} \; avec \; ici \sum_{i=1}^{10} f_{i/j} = 1 \; ou \sum_{i=1}^{n} f_{i/j} = 1$$

avec n = nombre de modalités de la variable X

*b) en ligne*

$$f_{j/i} = \frac{n_{i,j}}{n_i.} \; avec \; ici \sum_{j=1}^{2} f_{j/i} = 1 \; ou \sum_{i=1}^{p} f_{j/i} = 1$$

avec p = nombre de modalités de la variable Y

sept.-20 JF Léger - IDUP 13

Les fréquences conditionnelles correspondent au poids d'une catégorie d'individus définie au croisement des deux variables X et Y au sein d'une sous-population définie par une modalité de la variable X OU Y.

Par exemple, parmi les personnes ayant toutes en commun la modalité j de Y (condition première), la part de ceux qui sont caractérisés par la modalité i de X est donnée par la relation suivante :

$$f_{i/j} = \frac{n_{ij}}{n_{.j}}$$

Elle est notée $f_{i/j}$ ($f_i$ sachant j). La population de référence est donc la population caractérisée par la modalité j de Y, soit $n_{.j}$. La fraction de population dont on cherche à calculer le poids au sein de cette sous-population est $n_{ij}$. Dans le calcul des proportions, l'effectif de référence (ici $n_{.j}$) est toujours placé au dénominateur. Celui correspondant à la population dont on cherche à déterminer la proportion (ici $n_{ij}$) est toujours placé au numérateur.

Par exemple $f_{3/2}$ correspond à la proportion de personnes caractérisées par la modalité 3 de X parmi les personnes caractérisées par la modalité 2 de Y. Dans le tableau relatif à la population lyonnaise, $f_{i=3/j=2}$ correspond à $f_{6\text{-}10\ ans/femme}$, c'est-à-dire à la proportion de personnes âgées de 6-10 ans au sein de la population féminine.

$f_{j/i}$ ($f_j$ sachant i) correspond à la proportion de personnes caractérisées par la modalité j de Y parmi celles caractérisées par la modalité i de X (condition première). Par exemple, dans la population lyonnaise, $f_{j=1/i=5}$ correspond à la proportion d'hommes (modalité 1 de Y) parmi les personnes âgées de 18-24 ans (modalité 5 de X). On peut noter cette proportion conditionnelle : $f_{Hommes/18\text{-}24\ ans}$

## Relations entre fréquences élémentaire, conditionnelle et marginale

a) Fréquence élémentaire = fréquence marginale × fréquence conditionnelle

$$\frac{n_{ij}}{n..} = \frac{n_{i.}}{n..} \times \frac{n_{ij}}{n_{i.}} \quad \text{ou} \quad \frac{n_{ij}}{n..} = \frac{n._j}{n..} \times \frac{n_{ij}}{n._j}$$

b) Fréquence conditionnelle = Fréquence élémentaire / Fréquence marginale

$$\frac{n_{ij}}{n_{i.}} = \frac{\frac{n_{ij}}{n..}}{\frac{n_{i.}}{n..}} = \frac{n_{ij}}{n..} \times \frac{n..}{n_{i.}} \quad \text{ou} \quad \frac{n_{ij}}{n._j} = \frac{\frac{n_{ij}}{n..}}{\frac{n._j}{n..}} = \frac{n_{ij}}{n..} \times \frac{n..}{n._j}$$

c) Fréquence marginale = Fréquence élémentaire / Fréquence conditionnelle

$$\frac{n_{i.}}{n..} = \frac{\frac{n_{ij}}{n..}}{\frac{n_{ij}}{n_{i.}}} = \frac{n_{ij}}{n..} \times \frac{n_{i.}}{n_{ij}} \quad \text{ou} \quad \frac{n._j}{n..} = \frac{\frac{n_{ij}}{n..}}{\frac{n_{ij}}{n._j}} = \frac{n_{ij}}{n..} \times \frac{n._j}{n_{ij}}$$

sept.-20 JF Léger - IDUP 14

**Ne pas retenir les formules par cœur** ! Il faut juste garder en tête qu'il existe des relations entre ces trois types de fréquences relatives et qu'il est toujours possible d'en déduire une à partir des deux autres.

Attention : cela est vrai seulement si les fréquences marginale et conditionnelle ont en commun :

- une modalité i de la variable X : proportion de personnes ayant la modalité i de X dans la population totale (fréquence marginale : part du **sous-groupe i de X**) et proportion de personnes ayant la modalité j de Y parmi ceux qui ont la modalité i de X (fréquence conditionnelle : part de j de Y dans le **sous-groupe i de X**) ;
- ou une modalité j de la variable Y : proportion de personnes ayant la modalité j de Y dans la population totale (fréquence marginale : part du **sous-groupe j de Y**) et proportion de personnes ayant la modalité i de X parmi ceux qui ont la modalité j de Y (fréquence conditionnelle : part de i de X dans le **sous-groupe j de Y**).

Par exemple :

Fréquence élémentaire = fréquence marginale × fréquence conditionnelle

C'est-à-dire : $f_{ij} = f_i \times f_{j/i}$ ou $f_{ij} = f_j \times f_{i/j}$

Pour des exemples, cf. les corrigés des exercices faits en cours.