Statistique descriptive en SHS

Chapitre 1 : Description d'une situation statistique et distribution

Département MIAp - UFR 6 - UPV-UM3

Licence 1re année

Sommaire

Sommaire

Les individus

Réponse à la question : “**sur qui** porte l’étude ?”

Pour désigner un individu, on parle aussi d’**unité statistique**.
Les individus ne sont pas nécessairement des “personnes”.

On distingue les individus que l’on a réellement observés, mesurés, interrogés ... de l’ensemble des individus concernés par l’étude. C’est la différence entre l’**échantillon** et la **population**.

La **totalité** des individus sur lesquels porte l’étude constitue **la population**.
Il est souvent **impossible** (ou au moins très peu pratique) d’étudier la population dans son ensemble. Dans ce cas, on se contente d’en extraire une **partie** (un sous-ensemble) que l’on appelle **échantillon**.
Les individus qui constituent l’échantillon sont donc extraits de la population étudiée.

Choisir les individus de l’échantillon **est tout un art** ! Il s’agit en effet que l’échantillon soit **représentatif** de la population.

Les individus

Le **nombre** d'individus qui composent l'échantillon est appelé **la taille** de l'échantillon.

Un échantillon constitue une vue nécessairement **partielle**, **approximative** de la population ... mais on espère bien que l'information qu'il porte nous permette de tirer des conclusions pour la population entière. Il s'agit alors **à partir de l'échantillon** d'**inférer** des **propriétés sur la population** : ce domaine constitue la **statistique inférentielle**, à différencier de la **statistique descriptive** qui se limite à décrire l'échantillon.

Dans les quelques cas où l'on a pu étudier l'ensemble de tous les individus de la population, on parle alors de **recensement**. L'échantillon correspond alors à la population entière.

Les variables

Réponse à la question : “**sur quoi** porte l’étude ?”

On parle de **caractère** ou **variable** que l’on observe, que l’on mesure sur chaque individu.

Une variable est désignée par une **lettre majuscule :** X**,** Y**,** U ...
L’observation qui en est faite **varie** d’un individu à l’autre.

On appelle **modalités** les réponses faites par les individus à une variable.
Un individu n’a **qu’une seule** réponse possible. Sa réponse est désignée par une **lettre minuscule**, par exemple x_3 désigne la réponse faite par l’individu numéro 3 de l’échantillon à la variable X.

On distingue l’ensemble des modalités **observées** de l’ensemble des modalités **observables**. Il est en effet possible qu’au travers des individus de l’échantillon toutes les réponses n’aient pas été rencontrées, soit parce que l’ensemble des modalités observables est infini, soit parce que l’échantillon n’a pas pu recouvrir l’ensemble des possibilités.

Les variables

On désignera par $\mathcal{U}_X$ l'ensemble des modalités de la variable X.
Par exemple, pour une variable X "choix d'une activité" pour les enfants, on a $\mathcal{U}_X = \{$lecture, sport, peinture, musique$\}$.

Dans le cas où l'ensemble des modalités est **fini**, on note C son **cardinal**. On a alors :

$$\mathcal{U} = \{m_1, m_2, ..., m_C\},$$

Lorsque $C = 2$, la variable est dite **dichotomique**.

Les variables

Il est fondamental pour la suite de l'analyse statistique de savoir étudier la **structure** de cet ensemble de modalités.

1. Certaines modalités sont des **mots** : la variable est dite **qualitative**. Les modalités sont aussi appelées des **niveaux** et on regarde s'il existe un **ordre naturel** sur ces modalités :
   - si non, variable qualitative **nominale**
   - si oui, variable qualitative **ordinale**

2. D'autres modalités sont des **nombres** (on a compté, mesuré ...) : la variable est dite **quantitative**. On parle alors de **valeurs** plutôt que de modalités :
   - si les valeurs sont isolées les unes des autres, variable quantitative **discrète**
   - si les valeurs sont prises dans des intervalles, variable quantitative **continue**

Les données

Réponse à la question : “**quel relevé** des observations ?”

On distingue les **données brutes** des données qui ont déjà été regroupées. Les **données brutes**, sous forme de **tableau** ou de **liste**, sont le **relevé complet en pratique de l'information**. Aucune opération n'a encore été réalisée sur les réponses.

Exemple :

• On interroge 10 employés d'une entreprise pour savoir dans quel service ils travaillent. Les réponses sont :
vente, logistique, vente, direction, production, logistique, production, production, gestion, production

• On interroge 100 employés d'une entreprise pour savoir dans quel service ils travaillent.

Service	production	logistique	vente	gestion	direction
Effectifs	66	14	8	7	5

Les données

Notations :

Dans tout ce cours, on notera désormais

$$\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$$

les n réponses des n individus de l'échantillon à la variable X.
$\hookrightarrow$ la liste des données brutes

et

$$\mathbf{x_{(1)}}, \mathbf{x_{(2)}}, \ldots, \mathbf{x_{(n)}}$$

les mêmes réponses mais ordonnées (par ordre croissant) pour les variables quantitatives uniquement.
$\hookrightarrow$ la liste des données brutes ordonnées

Sommaire

La distribution

La **distribution d'une variable** est la **répartition** des individus, selon leur réponse, sur les différentes modalités de la variable.

Exemple :
On interroge 100 employés d'une entreprise pour savoir dans quel service ils travaillent.

Service	production	logistique	vente	gestion	direction
Effectifs	66	14	8	7	5

À chaque **modalité** est donc associé un **effectif** correspondant au nombre d'individus ayant eu cette modalité pour réponse à la variable X.

La distribution

Dans le cas général, les effectifs seront notés n_k et **la distribution en effectif d'une variable qualitative X pourra être représentée par** :

Modalités m_k	m_1	m_2	$\ldots$	m_C	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n

La somme des effectifs n_k est égal à n : $\sum_{k=1}^{C} n_k = n$.

$\longrightarrow$ **La distribution en fréquence est donnée par l'ensemble des** $f_k = \frac{n_k}{n}$ associés aux modalités m_k :

Modalités m_k	m_1	m_2	$\ldots$	m_C	Total
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1

Attention : La somme des fréquences d'une distribution est égale à 1 (du fait de la répartition des individus sur toutes les modalités), soit $\sum_{k=1}^{C} f_k = 1$.

La distribution

Dans l'exemple, cela donne :

Service	production	logistique	vente	gestion	direction
Fréquences	0.66	0.14	0.08	0.07	0.05

Les mêmes tableaux pourront être utilisés pour les distributions d'une variable quantitative discrète (avec les valeurs v_k à la place des modalités m_k) et d'une variable quantitative continue (les classes $[b_{k-1}; b_k[$ jouant ici le rôle de modalités, voir le chapitre suivant).