Statistique descriptive

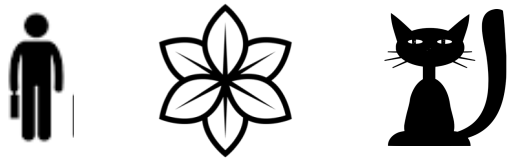
Description d'une situation statistique et distribution

Les ingrédients de la description

Les individus - “sur qui porte l’étude ?”

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

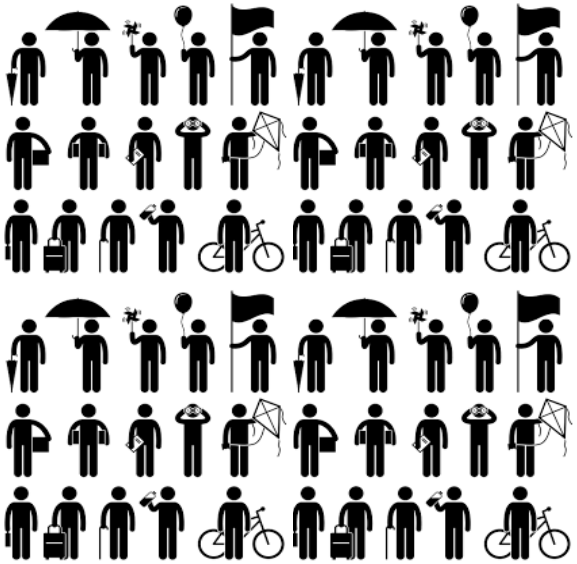
Individus



Les individus - “sur qui porte l’étude ?”

- Il est souvent **impossible** (ou peu pratique) d’étudier la population dans son ensemble.
- On se contente d’en extraire une **partie** (un sous-ensemble) que l’on appelle **échantillon**.
  - L’**échantillon** correspond aux individus que l’on a réellement observés, mesurés, interrogés...
  - La **population** correspond à l’ensemble des individus concernés par l’étude.

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



Population

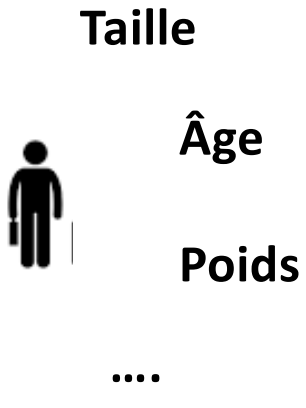
Echantillon de taille 15

Les individus- “sur qui porte l’étude ?”

- Choisir les individus de l’échantillon **est tout un art** ! Il s’agit en effet que l’échantillon soit **représentatif** de la population.
- Un échantillon est une vue **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.
- On distingue :
  - **statistique descriptive :** description de l’échantillon
  - **statistique inférentielle :** à partir de l’échantillon, inférer des propriétés sur la population
  - **recensement :** cas où l’on a pu étudier l’ensemble de tous les individus de la population (échantillon = population entière)

Les variables- “sur quoi porte l’étude ?”

- On parle de **caractère** ou **variable** que l’on observe, que l’on mesure pour chaque individu.
- Une variable est désignée par une **lettre majuscule** : X , Y , U ...
- L’observation de cette variable **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 par variable
  - 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



Les variables- “sur quoi porte l’étude ?”

- 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- “sur quoi porte l’étude ?”

- On désignera par U_X l’ensemble des modalités de la variable *X*.
- Dans le cas où l’ensemble des modalités est **fini**, on note *C* son **cardinal**.

$$U_X = \{m_1, m_2, \ldots, m_C\}$$

- Exemple : pour une variable *X* “choix d’une activité" pour les enfants, U_X = {lecture, sport, peinture, musique}.

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

Les variables- “sur quoi porte l’étude ?”

- **Variable qualitative :** les modalités sont des **mots**
  - **nominale :** il n’existe pas d’ordre entre les modalités
    Exemple : U_X = {lecture, sport, peinture, musique}
  - **ordinale** : sinon
    Exemple : U_X = {Non satisfaisant, Peu satisfaisant, Satisfaisant, Très satisfaisant}

- **Variable quantitative :** les modalités sont des **nombres** (on a compté, mesuré ...). On parle alors de **valeurs** plutôt que de modalités :
  - **discrète :** si les valeurs sont isolées les unes des autres
    Exemple : lancé de dé U_X = {1, 2, 3, 4, 5, 6}
  - **continue :** si les valeurs sont prises dans des intervalles
    Exemple : le nombre de femmes qui ont voté lors des dernières élections

Les données- “quel relevé des observations ?”

- La **liste des données brutes** : $\mathbf{x_1}$, $\mathbf{x_2}$, ..., $\mathbf{x_n}$ les *n* réponses des *n* individus de l’échantillon à la variable *X*.
- La **liste des données brutes ordonnées** : $\mathbf{x_{(1)}}$, $\mathbf{x_{(2)}}$, ..., $\mathbf{x_{(n)}}$ les mêmes réponses mais ordonnées (par ordre croissant) pour les variables quantitatives uniquement.
- La **distribution d’une variable** : **répartition** des individus, selon leur réponse, sur les différentes modalités de la variable.
- À chaque **modalité** est associé un **effectif** correspondant au nombre d’individus ayant eu cette modalité pour réponse à la variable *X*.
- 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

La distribution

- Les effectifs seront notés n_k
- La **distribution en effectif d'une variable qualitative** *X* est représentée par :

Modalités m_k	m_1	m_2	...	m_C	Total
Effectifs n_k	n_1	n_2	...	n_C	n

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

- La **distribution en fréquence** est donnée par l'ensemble des : $f_k = n_k/n$

Modalités m_k	m_1	m_2	...	m_C	Total
Fréquences f_k	f_1	f_2	...	f_C	1

La somme des fréquences d'une distribution est égale à 1

La distribution

- La somme des effectifs n_k est égal à n :

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)
  - d'une variable quantitative continue (les classes $[b_k - 1 ; b_k]$ jouant le rôle de modalités).