

ÉLÉMENTS DE TERMINOLOGIE

4A1ACD Analyse de données



LE SONDAGE

LE SONDAGE

- Pour réaliser un sondage, on crée un questionnaire avec un ensemble de questions
 - Certaines questions sont *fermées*, le choix des réponses est fixé à l'avance
 - Certaines questions sont *ouvertes*, les réponses seront différentes pour chaque personne (potentiellement)
 - Certaines questions portent sur des mesures *chiffrées* (combien d'employés y a-t-il dans votre entreprise? Quel âge avez-vous? Combien avez-vous payé pour cet ordinateur?)
 - Certaines questions portent sur des *opinions*
 - Certaines questions portent sur des *catégories*
- Le questionnaire est ensuite administré à l'échantillon

LE SONDAGE

- L'objectif du sondage est d'obtenir une réponse pour chacune des questions de chaque unité de sondage*
- Chaque **question** donnera lieu à une **variable** dans la base de données
- Les différentes réponses obtenues pour chacune des variables sont appelées des **modalités**

*: l'unité de sondage est le terme générique, il peut s'agir d'un consommateur, d'une personne représentant le point de vue d'une entreprise, etc.

LE SONDAGE

- Le questionnaire va être administré à un échantillon de la population
- Chaque personne ayant complété le questionnaire entraîne la création d'une chaîne de réponses (oui, bleu, panzani, 27, femme, ...)
- Cette chaîne de réponses constitue une **observation**

LA BASE DE DONNÉES

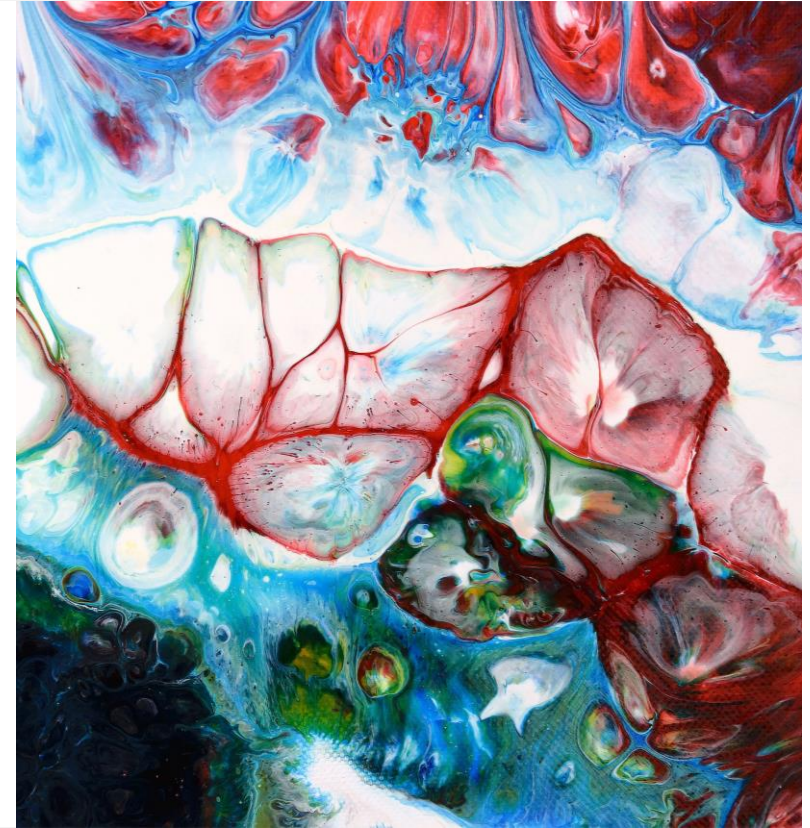
LA BASE DE DONNÉES (BDD)

- L'ensemble des réponses obtenues pour l'ensemble des personnes interrogées constitue la **base de données** (ou les données brutes)
- La BDD est une **matrice** où les lignes sont les observations et les colonnes sont les variables
 - une ligne correspond aux réponses d'une personne
 - une colonne correspond aux réponses de toutes les personnes de l'échantillon à une question
- La matrice sera constituée de **$k * j$ cellules**, où
 - k =nombre de colonnes, soit le *nombre de variables* et
 - j =nombre de lignes, soit la *taille de l'échantillon*

VARIABLE

Par définition, une variable est une:

- *Application définie sur une population statistique, qui associe une **valeur** numérique ou pas, à **chaque individu**.*
- Par exemple, à chaque individu interrogé, correspond une valeur telle que **l'âge**
- Les différentes valeurs prises par la variable **âge** sont les **modalités**



EXEMPLE

Lisez le tableau et répondez aux questions

Pays de résidence	Nombre de touristes (en 000)
Europe	69 292
Amériques	6 462
Asie, Océanie	5 213
Afrique	2 076
Total	83 013

- Quelle est la population étudiée ?
- Quelle est la variable ?
- Combien de modalités ?



NATURE DES VARIABLES

*La nature des variables
influence les analyses
qu'on peut en faire*

COMPARAISON

On fait généralement la différence entre des variables numériques et des variables qualitatives

Numérique

- Une variable est dite *numérique* si ses modalités sont des *nombres*
- Cette variable correspond à des questions du type:
 - Quel est votre âge ?
 - Combien avez-vous payé pour ce yaourt ?
 - Combien d'enfants avez-vous ?
 - Combien de pièces y a-t-il dans votre logement ?
- Pour chacune de ces questions, les modalités sont représentées par des nombres

Qualitative

- Une variable est dite qualitative si ses modalités ne sont *pas* des nombres
- Les modalités peuvent être des mots, des symboles, des codes, des échelles, etc.
 - Quelle est la marque de votre voiture ?
 - Quel est votre sexe ?
 - Quelle est votre lieu de résidence ?
 - Dans quelle mesure êtes vous d'accord avec l'affirmation suivante : ...
- Attention: un code postal, une échelle de mesure, etc. peuvent être des chiffres, mais il ne s'agit *pas* d'une valeur *numérique*

VARIABLES NUMÉRIQUES

Il existe deux catégories principales de variables numériques

Variable numérique continue

- Une variable numérique est *continue* si la valeur de ses modalités relève d'une mesure; elle peut prendre toutes les valeurs entre $-\infty$ et $+\infty$
- Une variable numérique continue peut être exprimée en *décimales*
 - Le prix de l'essence, la taille d'un individu, le salaire voire l'âge
- Cela ne signifie pas pour autant qu'on l'exprimera toujours en décimales

Variable numérique discrète

- Une variable numérique est *discrète* si la valeur de ses modalités relève d'un comptage
- Une variable numérique discrète est toujours un *nombre entier*
 - Nombre d'enfants, nombre de voitures, ...
- Les règles de traitement statistique sont les mêmes que pour des variables numériques continues
- Pour savoir si une variable numérique est continue ou discrète, il faut se poser la question si une décimale a du sens (2,3 enfants ?)

VARIABLES QUALITATIVES

Il existe deux grandes catégories de variables qualitatives (les modalités ne sont pas des nombres)

Variables catégoriques

- Une variable est dite **catégorique** (ou *nominale*) lorsque le but de la question est de catégoriser les réponses
- Les modalités d'une variable catégorique n'ont pas d'ordre spécifique (autre que alphabétique)
- Il s'agit simplement de catégories
 - Quel est votre sexe ?
 - Homme/Femme
 - Quelle est la marque de votre voiture ?
 - Audi, BMW, Citroën, ... Renault, etc.
 - Quelle était la raison principale d'achat ?
 - Confort, design, prix, consommation, ...

Variables ordinales

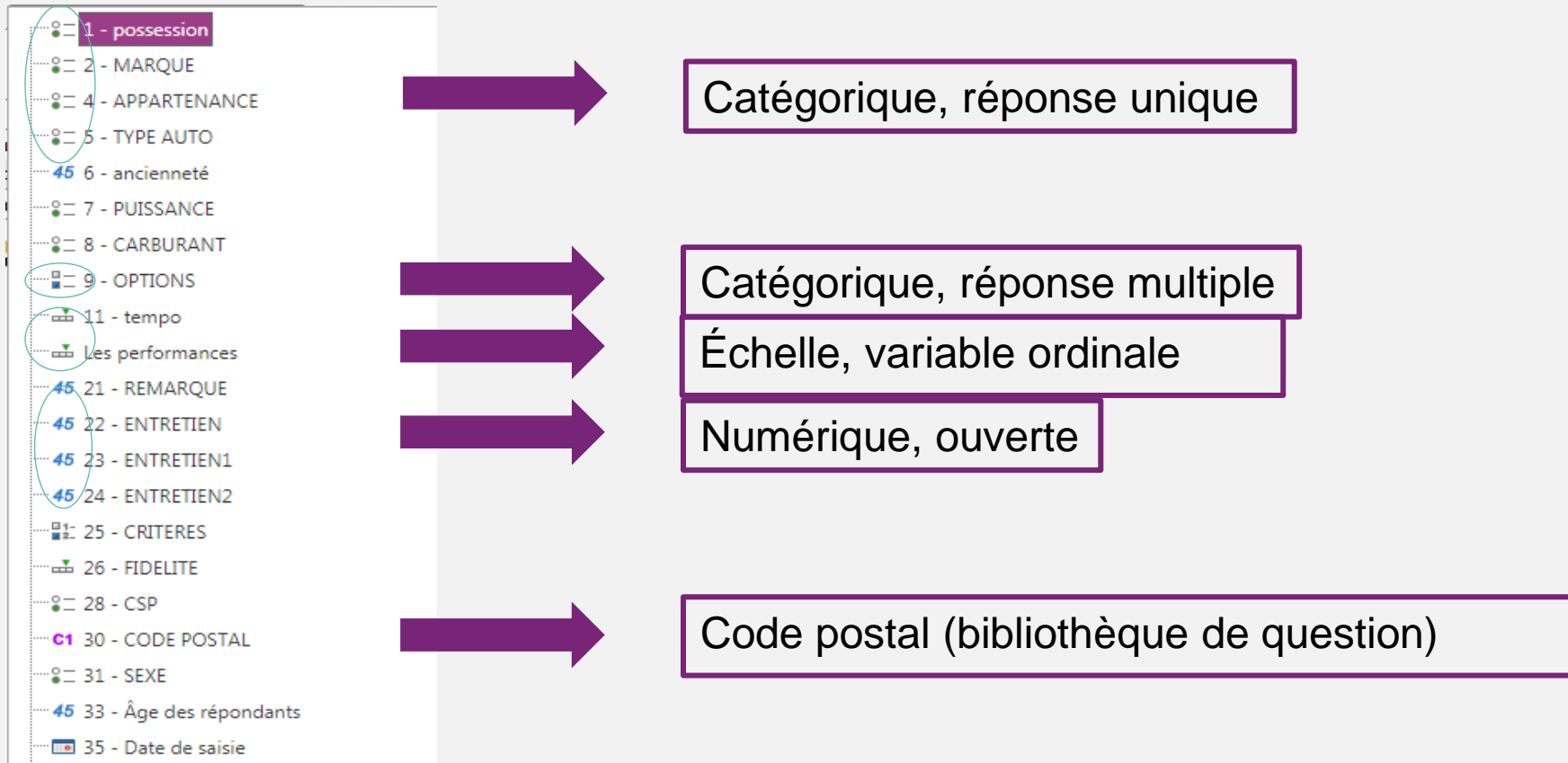
- Inversement, une variable est dite **ordinaire** lorsqu'il est possible de classer ses modalités dans un ordre logique
 - Dans quel groupe d'âge vous situez-vous ?
 - -18 ans; 18-24 ans; 25-34 ans, ...
 - Sur une échelle de *pas du tout probable* à *très probable*, indiquez dans quelle mesure vous êtes susceptible d'acheter ABC au cours des 6 prochains mois ?
 - Veuillez indiquer votre degré d'accord sur une échelle de 1 à 5, où 1 indique *pas du tout d'accord* et 5 indique *tout à fait d'accord*
 - *Attention! Les modalités 1, 2, 3, 4 et 5 ne sont pas considérées comme des variables numériques.*

ASTUCE POUR IDENTIFIER LA NATURE DE LA VARIABLE DANS SPHINX

Dans Sphinx Campus, la liste des variables est précédée d'une *icône* indiquant sa nature (Module « Analyse »)

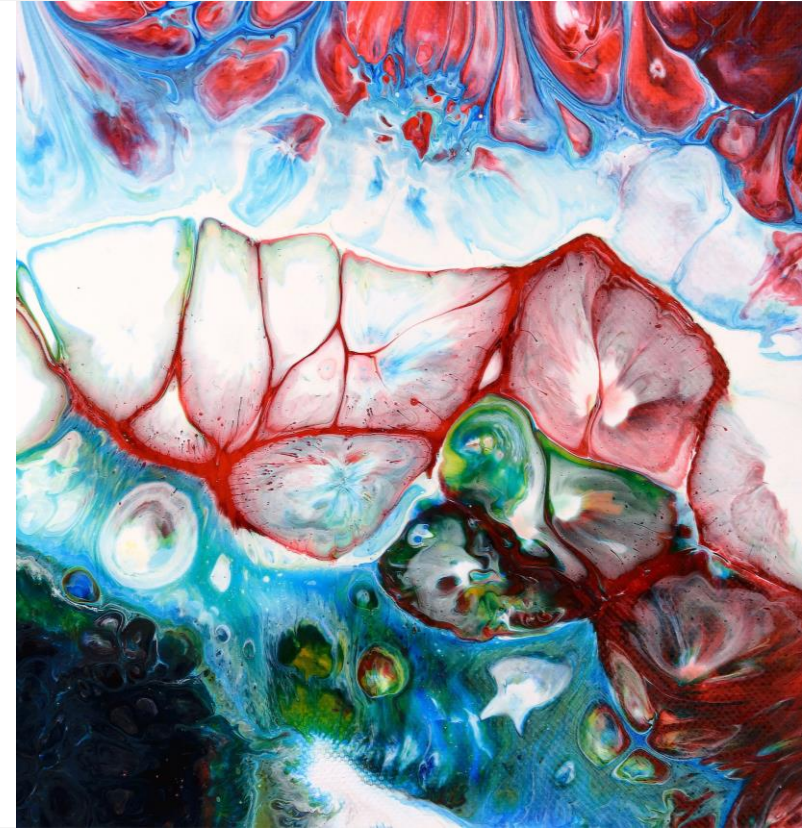
	Effectifs	% Obs.
Renault	36	20.6%

SIGNIFICATION DES ICÔNES



EFFECTIFS

- Effectif (ou fréquence absolue) :
 - **Nombre** d'individus statistiques présentant la modalité x_i
- Il s'agit de la **somme** de tous les individus ayant répondu à cette modalité
- Il y a **36** propriétaires de Renault, réponse à la question 2



FRÉQUENCE RELATIVE

- **Proportion** d'individus dans l'échantillon présentant la modalité x_i
 - On divise l'effectif par la taille de l'échantillon
 - La fréquence relative peut s'exprimer en décimale ou en pourcentage
- Il y a 20,6% des répondants qui sont propriétaires de Renault, réponse à la question 2 (36/175)



GÉRER LA BASE DE DONNÉES SUR SPHINX

*Veillez ouvrir
l'enquête
AUTOMOBILES*