Séance 1 : les bases de données

Introduction à la sociologie quantitative, niveau 1

Samuel Coavoux

Notions fondamentales des données

Données

Donnée : information sur un individu.

Individu / **observation** : unité statistique fondamentale. Entité à propos de laquelle on collecte des informations.

Variable : série de données collectée sur un ensemble d'individu, renseignant la même information (parfois appelé **caractère**).

Base de donnée

Série de variables à propos des mêmes individus.

Une base de donnée est un tableau contenant :

- en ligne, les observations (une observation par ligne)
- en colonne, les variables (une variable par colonne)
- dans chaque cellule, la **valeur** que prend une variable pour une observation

Base de donnée

nom	sexe	age
Dominique	homme	50
Camille	femme	20
Marie	femme	40

Population et échantillon

Population (ou population-mère) : ensemble des individus que l'on souhaite connaître.

Échantillon : ensemble des individus appartenant à la population à propos desquels on produit des données.

L'individu statistique

Individu au sens statistique est une entité, mais *pas nécessairement* une personne – d'où le fait que *observation* est un terme plus adapté que *individu*.

Autres possibilités fréquentes

- un ménage
- un pays, une région
- une institution (établissement, entreprise)
- un événement

Mais aussi plus généralement une unité d'observation

- un texte (si la population est un corpus)
- un objet (si la population est un ensemble d'objets)

L'individu statistique

Cette base de donnée recense les subventions à des projets culturels octroyées par l'état de New York.

Grant.Year	Program	Project.Title
2015	Music	Bach & Beyond Baroque Music Festival
2015	Theatre	General Operating Expenses
2015	Folk Arts	Traditional Czech Puppet Plays
2015	Theatre	A season of Theatre plays with Marionett
2015	Dance	92nd St Y Dance Space Grant Program : 2,0
2015	Literature	92Y Unterberg Poetry Center Main Reading

Données brutes et données agrégées

Données brutes : séries de valeurs associées à des individus. Informations *telles qu'elles sont recueillies*.

Données agrégées : chiffres produits par une analyse de donnée.

Dans le langage courant, on utilise souvent le terme "donnée" dans le second sens.

Exemple : un sondage mesure les intentions de vote d'un échantillon d'électeurs lors de la prochaine élection. On qualifiera de *donnée* l'affirmation "15 % des électeurs interrogés affirment vouloir voter pour le candidat X." Il s'agit d'une donnée agrégée.

Changements d'échelles

Il est parfois possible de changer de population par l'agrégation de données. On change alors la nature de l'individu.

state	revenu	feminisation
(01) Alabama/AL	34,894.78	54.11%
(02) Alaska/AK	34,662.70	53.14%
(04) Arizona/AZ	38,419.76	47.73%
(05) Arkansas/AR	32,556.87	52.59%
(06) California/CA	56,487.54	44.42%
(08) Colorado/CO	42,857.05	51.05%
(09) Connecticut/CT	53,939.37	51.43%
(10) Delaware/DE	37,884.62	46.49%

Bases de données appareillées

Parfois, l'analyse statistique s'intéresse à plusieurs types d'individus statistiques ou d'entités. La façon la plus efficace de stocker et de manipuler ces données consistent en la production de multiples bases appareillées par des clés ou identifiants permettant de passer d'une base à l'autre.

Bases de données appareillées

Je fais une enquête dans plusieurs entreprises en recueillant des données sur les salariés et sur les entreprises. Je peux faire une seule base incluant toutes les variables, mais certains informations sont redondantes.

statut	age	entreprise	nb_employes
cadre	45	A et co.	100-249
ouvrier	25	A et co.	100-249
cadre	36	B SARL	50-99

Bases de données appareillées

Je préfère deux bases : une pour les salariés, une pour les entreprises, avec un identifiant des entreprises pour faire le lien

statut	age	id_entrp
cadre	45	en01
ouvrier	25	en01
cadre	36	en02

id_entrp	nom	nb_employes
en01	A et co.	100-249
en02	B SARL	50-99

Les types de variables

Variable qualitative / quantitative

Quelle est la nature de l'information récoltée ?

- Une mesure => **variable quantitative**
- Une catégorie => **variable qualitative**

Variables qualitatives

Également variable **nominale** ou variable **catégorielle**. Les valeurs sont des catégories.

Modalités = ensemble des valeurs possibles pour une variable qualitative.

Cas particulier :

variables qualitatives **ordonnées** (parfois opposées aux variables **catégorielles**) = les modalités peuvent être classées dans un ordre logique.

Variables quantitatives

Mesure numérique. Composée d'une **mesure** (un nombre) et d'une **unité**.

Variable quantitative **discrète** : nombre restreint de valeurs possibles.

Variable quantitative **continue** : nombre de valeurs possibles important.

Exercice : Identifier les variables

sexe	age	state	income	dipl
Female	24	(37) North Carolina/NC	7,600.00	(19) 1 or more years o
Female	42	(36) New York/NY	18,000.00	(22) Master's degree
Female	64	(41) Oregon/OR	72,100.00	(22) Master's degree
Female	66	(55) Wisconsin/WI	19,400.00	(18) Some college, but
Male	50	(18) Indiana/IN	60,000.00	(21) Bachelor's degree
Female	40	(37) North Carolina/NC	24,500.00	(19) 1 or more years o

Recodage

Définition du recodage

Modification des modalités et des valeurs d'une ou de plusieurs variables afin de faciliter leur analyse statistique.

Objectifs :

- modifier le type de la variable
- réduire le nombre de modalités
- se débarrasser de valeurs rares ou aberrantes
- combiner plusieurs variables

Quantitatif -> qualitatif

Création de classes de valeurs. Par exemple, recoder l'âge en classes d'âge.

sexe	age	classe_age
Female	24	16-25 ans
Female	42	35-50 ans
Female	64	50-65 ans
Female	66	65-80 ans
Male	50	35-50 ans
Female	40	35-50 ans
Female	45	35-50 ans
Male	52	50-65 ans

Réduction du nombre de modalités d'une variable qualitative

Agréger des classes.

Objectifs :

- accroître la lisibilité
- améliorer les résultats des tests statistiques
- améliorer la qualité de description des classes

Réduction du nombre de modalités

Soit une variable avec un grand nombre de modalités

	n	%
(01) No schooling completed	291	0.2
(02) Nursery school, preschool	6	0.0
(03) Kindergarten	6	0.0
(04) Grade 1	10	0.0
(05) Grade 2	15	0.0
(06) Grade 3	36	0.0
(07) Grade 4	29	0.0
(08) Grade 5	41	0.0
(09) Grade 6	201	0.2
(10) Grade 7	84	0.1
(11) Grade 8	242	0.2
(12) Grade 9	462	0.4
(13) Grade 10	861	0.7
(14) Grade 11	1270	1.0

Réduction du nombre de modalités

Soit une variable avec un grand nombre de modalités

	n	%
(15) 12th grade - no diploma	822	0.7
(16) Regular high school diploma	12215	9.8
(17) GED or alternative credential	1680	1.4
(18) Some college, but less than 1 year	5996	4.8
(19) 1 or more years of college credit, no degree	18956	15.3
(20) Associate's degree	10745	8.7
(21) Bachelor's degree	50756	40.9
(22) Master's degree	15524	12.5
(23) Professional degree beyond a bachelor's degree	2268	1.8
(24) Doctorate degree	1507	1.2

Réduction du nombre de modalités

	n	%
Moins que bac	4376	3.5
Bac, début d'études supérieures	38847	31.3
Licence	61501	49.6
Master et plus	19299	15.6

Combinaison : l'accolade

Combinaison de deux variables qualitatives.

sexe	classe_age	age_sexe
Male	50-65 ans	Male 50-65 ans
Male	50-65 ans	Male 50-65 ans
Male	65-80 ans	Male 65-80 ans
Male	35-50 ans	Male 35-50 ans
Male	16-25 ans	Male 16-25 ans
Male	16-25 ans	Male 16-25 ans

Les sources de données

Les données ne sont pas données

Construire une bonne base de données est **très chronophage**.

- **Récupération des données**
  - Bases publiques (INSEE, INED, CEREQ) ou privées
  - Questionnaires ad hoc (en ligne ou papier)
  - Quantification d'archives
  - *Webcrawling* (Le Bon Coin), API (Twitter), etc.
  - Observation quantifiée
- **Recodage** (uniformiser les modalités)
- ***Sanity checks*** (traitement des incohérences dans les réponses)

Le questionnaire

Source de données la plus fréquente, qu'il s'agisse des bases publiques propices à l'analyse secondaire, ou de questionnaires *ad hoc* construits pour les besoins d'une enquête.

- Analyse secondaire : en France, la plupart des bases de données accessibles aux chercheurs en sciences sociales sont accessibles par le biais de l'ADISP et du réseau Quételet : http://www.reseau-quetelet.cnrs.fr (fournisseurs : INSEE, INED, Ministères)
- Questionnaires *ad hoc* : quelles règles de construction ?

Questionnaires *ad hoc*

- Hypothèses sociolgoiques
- Délimitation de la population
- Délimitation de l'échantillon
- Conception
- Passation
- Codage, nettoyage, analyse

Questionnaires *ad hoc* : population et échantillon

Le mode d'échantillonage permet de s'assurer que l'on soit capable de calculer la probabilité d'erreur lors des inférences statistiques. Avoir un échantillonage adapté est la condition *sine qua non* de l'inférence : sans cela, on ne peut absolument pas tirer de conclusions à partir de notre échantillon.

Ainsi, **tous** les tests d'hypothèses partent du principe que **l'échantillonage est aléatoire**.

Questionnaires *ad hoc* : échantillonnage aléatoire

- on connaît précisément les limites de la population (= on dispose d'un annuaire comportant l'ensemble des individus, ou l'on est en mesure de le générer)
- chaque individu a exactement la même chance d'être inclu dans l'échantillon que tous les autres

Exemple : on a une liste intégrale des élèves d'un lycée, on note chacun de leur nom sur un papier de taille égale, que l'on mélange parfaitement dans un grand chapeau, et on en tire cent au hasard.

Questionnaires *ad hoc* : échantillonnage non aléatoire

Exemple d'échantillonage **non-aléatoire** : on interroge des gens dans la rue "au hasard" (selon l'heure, la rue, etc., ce sont des personnes différentes ; et comme on ne peut arrêter tout le monde, sans règle stricte pour choisir qui arrêter, les enquêteurs vont tendre à privilégier les cibles "faciles" [proximité sociale ; compétence linguistique ; bonne volonté sur le thème du questionnaire ; etc.])

Questionnaires *ad hoc* : échantillonnage stratifié

Les échantillonages stratifiés sont considérés comme de bonnes approximations d'échantillonages aléatoires, lorsque ceux-ci sont impossibles.

Il est souvent impossible d'avoir un accès exhaustif à la population et de s'assurer de l'équiprobabilité des tirages ; *mais* il est quasiment toujours possible d'obtenir un annuaire de *grappes* de population, puis un annuaire des individus dans ces *grappes*.

Par exemple, il est difficile d'avoir la liste des logements à Lyon ; mais on peut faire la liste des rues ; et pour chaque rue, il est aisé de trouver le nombre de logements

Questionnaires *ad hoc* : échantillonnage stratifié

Un échantillonage stratifié consiste à tirer au sort, aléatoirement, des grappes ; puis à tirer aléatoirement des individus dans ces grappes (parfois, tirer des grappes, puis des sous-grappes, puis des individus).

Exemple : tirer au sort 10 quartiers d'une ville, puis 3 rues de chaque quartier, puis 1 ménage tous les 10 numéros de la rue (si plusieurs ménages dans le numéro : faire la liste et tirer au sort).

Questionnaires *ad hoc* : sur-représentation et pondération

Il est parfois nécessaire, dans l'échantillonage, de sur-représenter certaines populations afin d'atteindre un seuil minimal en-dessous duquel on ne peut rien dire.

Il y a environ 2% d'agriculteurs dans la population active française ; sur 1000 enquêtés, étant donné les inactifs, on peut donc s'attendre à en trouver une dizaine. C'est trop peu pour dire des choses de cette sous-population $=>$ si l'on souhaite pouvoir parler des agriculteurs sans pour autant augmenter le nombre d'enquêtés, on en interroge plus que leur proportion dans la population, mais on attribue à chacun un poids plus faible dans le total.

Questionnaires *ad hoc* : échantillonnage non-aléatoire

- Par quota : on fixe a priori des critères, en se basant sur une population connue. On interroge n'importe qui dans les proportions fixées par ces critères. Facile, mais dangereux (méthode très biaisée)
- Échantillon volontaire : enquêté auto-sélectionné (par exemple enquête sur Internet)
- Échantillon accidentel : on interroge n'importe qui.

Dans ce cas, l'usage des tests et des techniques statistiques en général est incorrect. La fameuse "marge d'erreur" des sondages est... tout simplement fausse ! Elle fait comme si l'échantillonage par quota était exactement équivalent à l'échantillonage aléatoire.

Questionnaires *ad hoc* : conception

Contraintes : temps et longueur

- contraintes matérielles, prix du questionnaire
- contraintes cognitives, attention du répondant

Ordre des questions important :

- filtres
- enchaînement des questions
- effet de la thématique sur les réactions de l'enquêté

Questionnaires *ad hoc* : conception

Importance de la rédaction de la question :

- neutre (pas de suggestions, rappels, etc.)
- simple (vocabulaire compréhensible)
- univoque
- doit soit avoir déjà été posée par l'enquêtée, soit pouvoir provoquer une réponse rapide, évidente.

Questionnaires *ad hoc* : conception

Types de questions :

- choix unique
- choix multiple
- ouverte (à éviter)

Questions fermées

- réponses exhaustive (toutes les réponses possibles)
- mais pas trop dispersées

Questionnaires *ad hoc* : passation

- auto-administré (papier/Internet)
- téléphone
- face à face

Plusieurs effets :

- économique
- cognitif
- taux de réponse
- relances

Passage du questionnaire à la base de données

En sciences sociales, de nombreuses bases de données sont produites par questionnaire (statistique publique, etc.).

En règle générale :

- un questionnaire = un individu
- une question = une variable

Mais ce n'est pas toujours le cas :

- questionnaires collectifs / multi-niveaux
- questions à choix multiples
- questions ouvertes
- filtres

Questionnaires collectifs ou multi-niveaux

Recensement :

Une personne du ménage remplit le questionnaire. Deux bases de données sont produites :

- une base des ménages (un questionnaire = un individu ; individu = ménage)
- une base des personnes (un questionnaire = autant d'individu que de membre du ménage ; individu = une personne)

Questions à choix multiples

Les questions à choix multiples apparaissent dans les bases de données comme autant de variables qu'il existe de choix possibles.

> *Nous allons maintenant parler de vos sorties. Je vais vous montrer une liste, dîtes-moi celles qu'il vous arrive de faire le SOIR, que ce soit le soir en semaine ou le soir en week-end. (Source : Enquête pratiques culturelles des Français, 2008)*

- Aller au cinéma
- Aller au spectacle
- Aller chez des parents
- Aller chez des amis
- Aller à une réunion autre que familiale ou amicale
- Aller au restaurant
- Aller vous promener, retrouver des amis dans la rue, au café...

Questions à choix multiples

cinema	spectacle	parents	amis	reunion	restaurant	promenade
non	non	oui	oui	oui	non	non
non	non	oui	oui	oui	non	non
oui	non	non	non	non	non	non
non	non	oui	oui	non	oui	non
non	non	non	non	non	non	non
oui	oui	oui	oui	non	oui	oui
non	oui	non	oui	non	oui	oui
non	non	non	non	non	non	non

Les “questions ouvertes”

Peuvent apparaître dans les bases de données dans une colonne, mais ne constituent pas une variable. Il est nécessaire de les nettoyer.

Les valeurs manquantes

Plusieurs raisons :

- la variable ne s'applique pas à l'individu (profession du conjoint pour une personne célibataire)
- l'information est indisponible pour un individu (refus de répondre, échec des recherches)

Souvent "codées" comme des réponses, et différenciées (nsp, refus de répondre, manquant). Dans l'analyse, choix nécessaire : les inclure, les ignorer, les transformer.

Bases de données d'observations ou d'archives

Les **mêmes règles** de délimitation de la population et de l'échantillon existent que pour les questionnaires. On a parfois accès à la population entière ; on a souvent la possibilité de créer un annuaire et de faire un tirage effectivement aléatoire.

Utiliser la possibilité du tirage : rien ne sert d'avoir une population totale si un échantillon peut suffire.

Bases de données d'observations ou d'archives

S'interroger sur **l'unité d'observation** la plus pertinente : l'individu, l'événement, la publication, etc.

Il est possible de créer plusieurs bases de données différentes et **appareillées** (susceptibles d'être rapprochées les unes des autres). Il importe alors de penser aux liens entre les bases de données (identifiants).

Bases de données d'observations ou d'archives

Claire Lemercier et Claire Zalc, *Méthodes quantitatives pour l'historien*, La découverte, 2008.

- Séparer la **saisie** du **codage**. (séparer les moments, voire les logiciels)
- Rester au plus près de la source lors de la saisie.
- Créer le plus de variables possibles