Méthodes quantitatives

Paul Hobeika

2022-01-28

Contents

Cette page accueille les notes de cours de méthodes quantitatives du M1 de Science Po Strasbourg.

Chapter 1

Données et vocabulaire de la statistique

1.1 Les sources statistiques en sociologie

Nous avons évoqué la semaine dernière l'importance de la connaissance des sources statistiques pour la production de savoirs quantitatifs en sciences sociale. Il en existe différents types, qu'il est important de savoir identifier.

1.1.1 Les enquêtes par questionnaire produites par les chercheur-es

C'est par exemple le cas des données exploitées dans *La distinction* [Bourdieu, 1979] dont nous avons parlé au premier semestre. A partir d'une problématique de départ parfois abstraite (dans le sens pas directement quantifiable), l'élaboration d'un questionnaire a souvent pour objectif de trouver des éléments empiriques concrets qui permettent de rendre opérationnelles certaines notions ou concepts. Par exemple, dans *La distinction*, le questionnaire porte sur les pratiques culturelles et permet d'opérationaliser empiriquement la notion de *capital culturel*.

> Remarque : si vous souhaitez produire vous-même des données dans le cadre de votre TER et de la validation du cours c'est tout à fait possible, mais nous n'aborderons pas la méthodologie du questionnaire dans ce cours. De bons manuels sont toutefois disponibles, je vous recommande par exemple celui de Bugeja-Bloch and Couto [2021], chapitres 3 et 4.

1.1.2 Les autres source de "première main"

En réalité, les chercheur-es peuvent effectuer des traitement quantitatifs sur d'autres types de sources que les données issues d'un questionnaire. Pour cette raison, Fanny Bugeja-Bloch et Marie-Paule Couto font une distinction entre les **technique d'enquête** et les **technique d'analyse** des données.

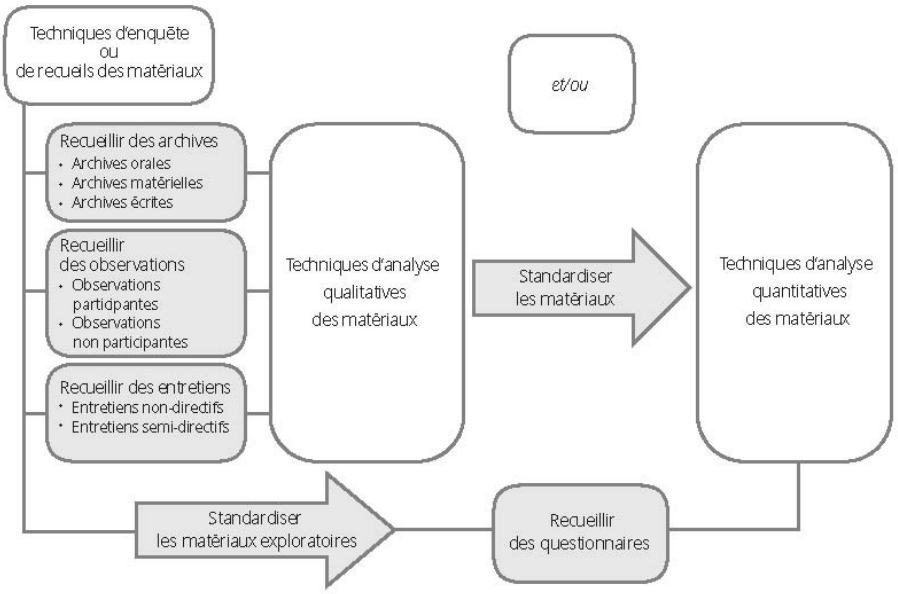

Figure 1.1: Techniques d'enquête et techniques d'analyse [Bugeja-Bloch and Couto, 2021]

Les **techniques d'enquête** désignent les différents modes de recueil des données : données d'archives, d'entretien ou encore issues d'observation. Les matériaux ainsi produits peuvent ensuite se prêter à différentes formes d'analyse. C'est seulement à ce niveau que l'on peut distinguer méthodes qualitatives et quantitatives. Une analyse qui se fondera sur le commentaire d'un ou quelques extrait d'entretien par exemple sera qualifiée de *qualitative*. Mais ces mêmes matériaux, lorsqu'ils sont *standardisés* et *mis en série* peuvent également être l'objet de techniques d'analyse quantitative. On peut produire des statistiques à partir d'archives [Lemercier and Zalc, 2008], à partir d'entretiens (le questionnaire en est un cas particulier) ou encore à partir d'observations [1].

[1] Un exemple tiré de la sociologie du travail est celui de l'enquête de Jean Peneff sur les urgences. Effectuant une enquête par observation participante en tant que brancardier dans un service d'urgence, il fait un certain nombre de comptages dans l'objectif d'objectiver certaines dimensions du travail aux urgences [Peneff, 1992].

1.1.3 L'analyse secondaire des données

Dans de nombreux cas, ce ne sont pas les sociologues ou politistes qui produisent les données qu'ils ou elles exploitent. On parle alors d'**analyse secondaire des données**. C'est le cas lorsqu'on travaille sur des données de l'Insee ou n'importe quelle base de donnée produite par une administration.

Quelques liens pour accéder aux données de la statistique publique française :

- le site de l'Adisp (Archives de données issues de la statistique publique) , qui rassemble les données de l'Insee et des directions statistiques ministérielles (santé, travail, culture, etc.)
- les données de l'Ined (Institut national de la recherche démographique)

1.1.4 Données d'enquête et données de gestion

Parmi l'ensemble des données accessibles produites par la statistique publique, on distingue en général deux grandes catégories [Desrosières, 2005] . D'un côté les bases de données produites via une **enquête par questionnaire** comme évoqué plus haut : elles sont réalisées à partir d'un échantillonage au sein d'une population plus large (voir plus loin pour des définitions de ces termes), et comportent un grand nombre de variables, qui correspondent en général à des questions qui sont posées directement par des enquêteurs ou enquêtrices. De l'autre côté, certaines bases de données sont le **résultat du travail de gestion de certaines administrations** : par exemple, les employeurs effectuent chaque année ce qu'on appelle une "déclaration annuelle de données sociales", dans laquelle ils renseignent une série d'informations sur leurs différents salarié · es (parmi lesquelles leur salaire et leur profession). Ces "DADS" constituent un exemple de base de données administrative. Ils sont largement utilisée pour étudier les salaires. Ces bases de données sont intéressantes mais en général moins riches que les données d'enquête, car elles ne sont pas produites dans le but de produire de la connaissance. Je vous conseille d'éviter de choisir une telle base de données, car il est souvent plus difficile d'en tirer des résultats intéressants à moins de savoir exactement ce qu'on cherche.

1.2 Le vocabulaire de la statistique

1.2.1 Bases de données

Il est temps d'expliquer plus précisément ce qu'on entend par "base de données". En voilà un premier exemple, issu du package R `titanic` :

Une **base de données** se présente sous la forme d'un tableau. Les lignes décrivent les **individus** : ici ce sont des passagers, mais gardez en tête que la nature des individus peut être à peu près n'importe quoi (ça peut être des ménages, des villes, des bactéries, n'importe quoi). Chaque colonne apporte des

Table 1.1: Extrait de la base de données des passagers du Titanic

PassengerId	Survived	Age	Name
1	0	22	Braund, Mr. Owen Harris
2	1	38	Cumings, Mrs. John Bradley (Florence Briggs Thayer)
3	1	26	Heikkinen, Miss. Laina
4	1	35	Futrelle, Mrs. Jacques Heath (Lily May Peel)
5	0	35	Allen, Mr. William Henry
6	0	NA	Moran, Mr. James

Table 1.2: Extrait de la base USArrests

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

éléments permettant de caractériser les individus (leur nom, leur âge, etc.). On appelle ces caractéristiques des **variables**.

1.2.2 Un autre exemple

Dans cet exemple, les individus sont des États des États-Unis. Les variables correspondent à des taux d'arrestation par la police pour meurtre, agression et viol pour 10000 habitants en 1973, ainsi que le pourcentage de la population urbaine.

1.2.3 Données "tidy"

Dans R, on qualifie certaines base de données de "tidy". C'est la structure qu'on souhaite avoir en général. Ces bases de données ont un individu par ligne, une variable par colonne. Dans chaque case, on trouve la modalité d'une variable correspondant à l'individu décrit dans la ligne

1.2.4 Séries temporelles

On parle parfois de **série temporelle** lorsque une base de donnée concerne un même individu statistique à différents instants. On parle dans ce cas là plutôt d'**observations** que d'individus. En voilà un exemple : la base de données `beaver1` accessible dans R présentent la température corporelle d'un castor en fonction du temps.

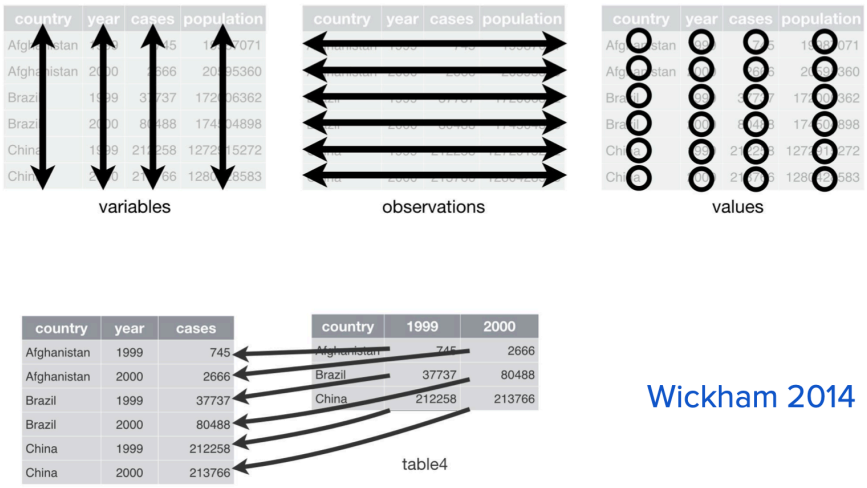

Figure 1.2: Tidy datasets

Table 1.3: Extrait de la base de données beaver1

day	time	temp	activ
346	840	36.33	0
346	850	36.34	0
346	900	36.35	0
346	910	36.42	0
346	920	36.55	0
346	930	36.69	0

1.2.5 Autres types de bases de données

- Des données qui décrivent différents individus à un moment donné sont parfois qualifiéed de **données en coupe** (ou *cross-sectional dataset*). Les données du titanic ou de USArrest en sont des exemples.
- Certaines bases de données décrivent un même groupe d'individus statistique de manière répétée dans le temps. On nomme ce genre de données des **données de panel** (exemple : enquête Emploi en continu).

1.3 Variables

1.3.1 Définition

Les variables sont les éléments qui permettent de décrire les individus présents dans le base de données. Lorsque les données sont issues d'un questionnaire, chaque question correspond en général à une variable.

Exemple :

- le sexe
- la catégorie socioprofessionnelle
- le niveau de diplôme
- le revenu

On appelle **modalités** les différentes valeurs que peuvent prendre une variable.

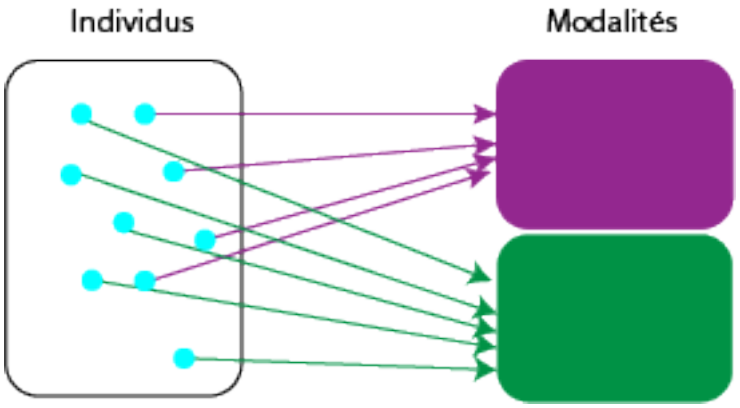

Figure 1.3: Les variables associent les individus à leur modalités

1.3.2 Variables qualitatives et variables quantitatives

On distingue les variables en fonction des opérations statistiques qu'on peut effectuer à partir de leurs modalités. Les deux grandes catégories de variables sont les **variables quantitatives**, dont les modalités sont des nombres, et les **variables qualitatives**, qui sont les autres.

1.3.3 Variables qualitatives

Parmi les variables qualitatives, on distingue encore :

- Les variables **qualitatives ordinales**, qui sont celles pour lesquelles ont peut ordonner les modalités (exemple : niveau de diplôme)
- Les variables **qualitatives nominales** dont les modalités ne sont pas hiérarchisables (exemple : sexe, catégorie socioprofesionnelle)

1.3.4 Variables quantitatives

Les modalités des variables quantitatives (ou numériques) sont des nombres qui ont une signification (par exemple, le code postal n'est pas une variable quantitative). Parmi elles, on distingue :

- Les variables **continues**, qui peuvent prendre toutes les valeurs réelles dans un intervalle donné
- Les variables **discrètes**, qui ne peuvent prendre qu'un certain nombre de valeurs

Pourquoi toutes ces catégories ? À ces différents types de variables, on associe différentes méthodes statistiques. Il est donc important de comprendre et mémoriser ces définitions, car lorsque vous souhaiterez étudier une variable, la première chose à faire sera d'identifier son type pour ensuite utiliser les méthodes statistiques appropriées.

1.4 Mesures de tendance centrale

Ce sont des manières de résumer l'information contenue dans une variable. En fonction du type de variable, il existe plusieurs indicateurs.

- La **moyenne** : lorsqu'on parle de moyenne, on fait généralement référence à la moyenne arithmétique d'un ensemble de valeurs numériques (par opposition à la moyenne géométrique ou harmonique). C'est une mesure très utilisée car elle fournit un premier résumé de la distribution statistique. Elle existe uniquement pour les **variables quantitatives**.

$$\bar{X} = \frac{1}{N} \sum_{i}^{N} x_i$$

- La **médiane** est la modalité d'une variable qui permet de séparer la population en deux parts égales. C'est-à-dire que 50% des individus auront une modalité supérieure ou égale à la médiane, et 50% une modalité inférieure ou égale à la médiane.

- Les **quantiles** sont une généralisation de la médiane : si vous voulez diviser votre population en groupe de 10%, vous pouvez utiliser les **déciles**. On appelle la médiane le **quantile d'ordre 2**, tandis que les déciles sont

les **quantiles d'ordre 10**. Les quantiles les plus utilisés sont la médiane, les quartiles, les déciles et les centiles.

- Les quantiles n'existent que pour les variables dont les modalités peuvent être hiérarchisées : toutes les variables quantitatives et les variables qualitatives ordinales.
- Le **mode** indique la modalité la plus fréquente d'une variable. Par exemple, la plupart des passagers du Titanic sont décédés dans le naufrage, donc le mode de la variable "Survived" est 0.

Le mode existe pour tous les types de variables.

Références

Bibliography

Pierre Bourdieu. *La distinction: critique sociale du jugement*. Les Editions de minuit, Paris, France, 1979. ISSN: 0768-049X.

Fanny Bugeja-Bloch and Marie-Paule Couto. *Les méthodes quantitatives*. Que sais-je ? PUF, 2021. OCLC: 1285669386.

Alain Desrosières. Décrire l'État ou explorer la société : les deux sources de la statistique publique. *Geneses*, no 58(1):4–27, 2005. URL https://www.cairn.info/journal-geneses-2005-1-page-4.htm. Bibliographie_available: 0 Cairn-domain: www.cairn.info Cite Par_available: 1 Publisher: Belin.

Claire Lemercier and Claire Zalc. *Méthodes quantitatives pour l'historien*. Number 507 in Repères. la Découverte, Paris, 2008.

Jean Peneff. *L'hôpital en urgence: étude par observation participante*. Métailié : Diffusion, Seuil, Paris, 1992.