Analyse statistique - Séance 2 : La catégorisation

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2025

Les statistiques ethniques

SIMON Patrick, « Les statistiques, les sciences sociales françaises et les rapports sociaux ethniques et de “race” », *Revue française de sociologie*, 2008, vol. 49, no 1, p. 153 162.

- Pourquoi créer des catégories statistiques sur la race ? Quels risques justifient l'opposition aux statistiques ethniques ou raciales ?
- Quelle voie a choisi la France ? Comment les chercheur·es travaillent sur ce rapport social ?
- Selon vous, les catégories statistiques doivent-elles s'inspirer des façons ordinaires dont les agents sociaux pensent et classent les individus ?

Le fait migratoire

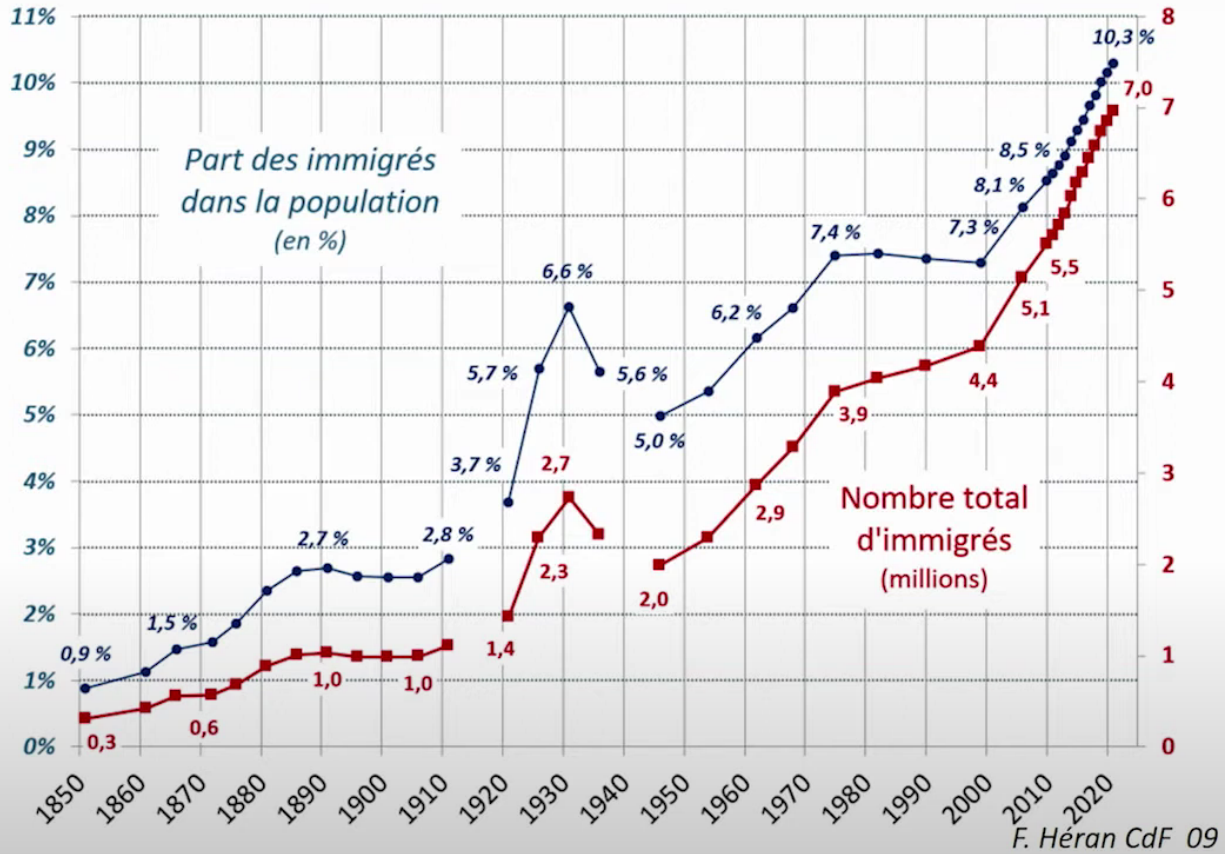
HERAN François, « Pour une vision historique et critique des droits des étrangers », 1er cours de la chaire *Migrations et sociétés*, année 2022-2023 : les migrations à la lumière du droit.

- Qu'est-ce que François Héran appelle le fait migratoire ?
- En quoi la fabrique du droit des étrangers est-elle liée à la mesure des flux démographiques ?

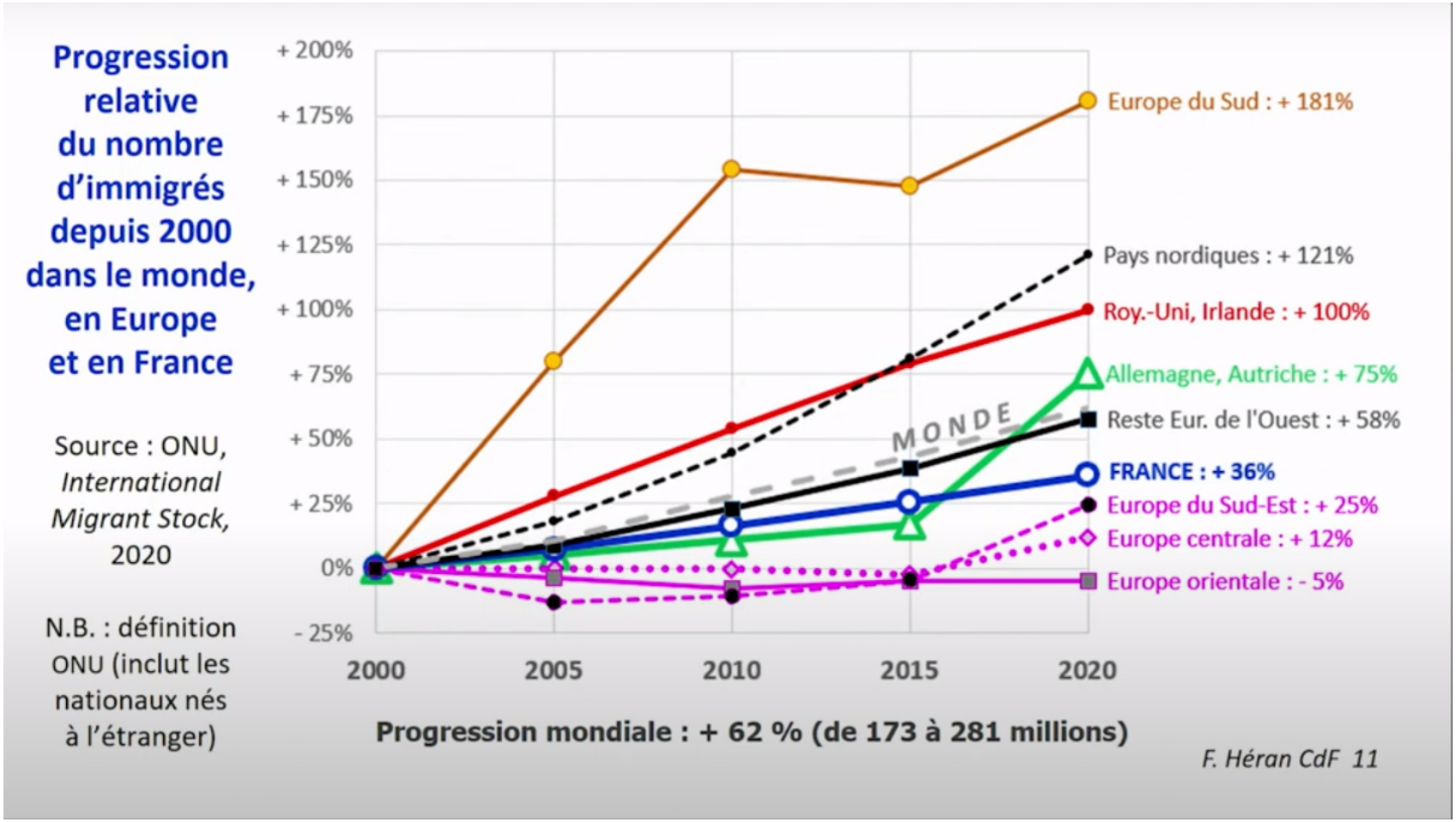
L'immigration en France depuis 1850

Nombre et proportion d'immigrés en France depuis 1850 : une hausse générale, affectée par les aléas de l'histoire économique et juridique

Source : recensements SGF et Insee



L'immigration dans le monde depuis 2000



Compter : définir le statut d'immigré·e

"Un immigré est une personne née étrangère à l'étranger et résidant en France.". Haut conseil à l'intégration, 1991.

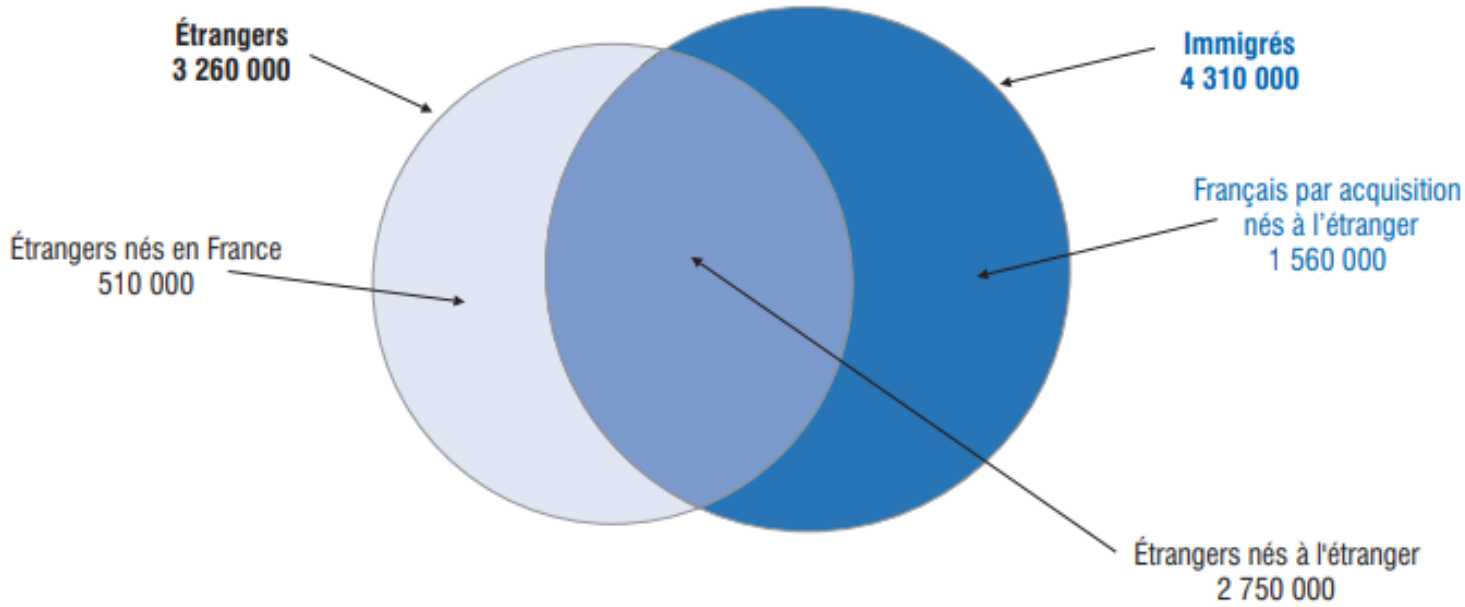
- Les personnes nées Françaises à l'étranger mais revenues en France ne sont pas immigrées.
- Les personnes nées étrangères devenues Françaises restent des immigrées.
- Les personnes nées étrangères en France ne sont pas immigrées.

Cette définition a elle-même connu des exceptions :

- Les personnes nées en Algérie avant 1962 étaient comptées comme étrangères nées à l'étranger même si elles disposaient de la nationalité Française ("musulmans algériens" dans le recensement de 1954 à 1962).
- Les personnes nées dans les territoires d'outre-mer ont été comptées comme immigrées dans certaines publications de la statistique publique.

Des catégories se recoupant partiellement

1 - Étrangers et immigrés



Source : Insee, Recensement de la population, 1999.

Compter : le bulletin individuel du recensement

Comment ces deux questions du recensement permettent-elles de compter le nombre d'immigré·es ?

2 Date et lieu de naissance

Né(e) le : jour mois année

à : commune (et arrondissement pour Paris, Lyon, Marseille)

département n° DOM pays pour l'étranger, territoire pour les COM

4 Quelle est votre nationalité ?

- **Française**
  - Vous êtes **né(e) français(e)** 1
  - Vous êtes **devenu(e) français(e)** (par exemple : par naturalisation, par déclaration, à votre majorité) 2

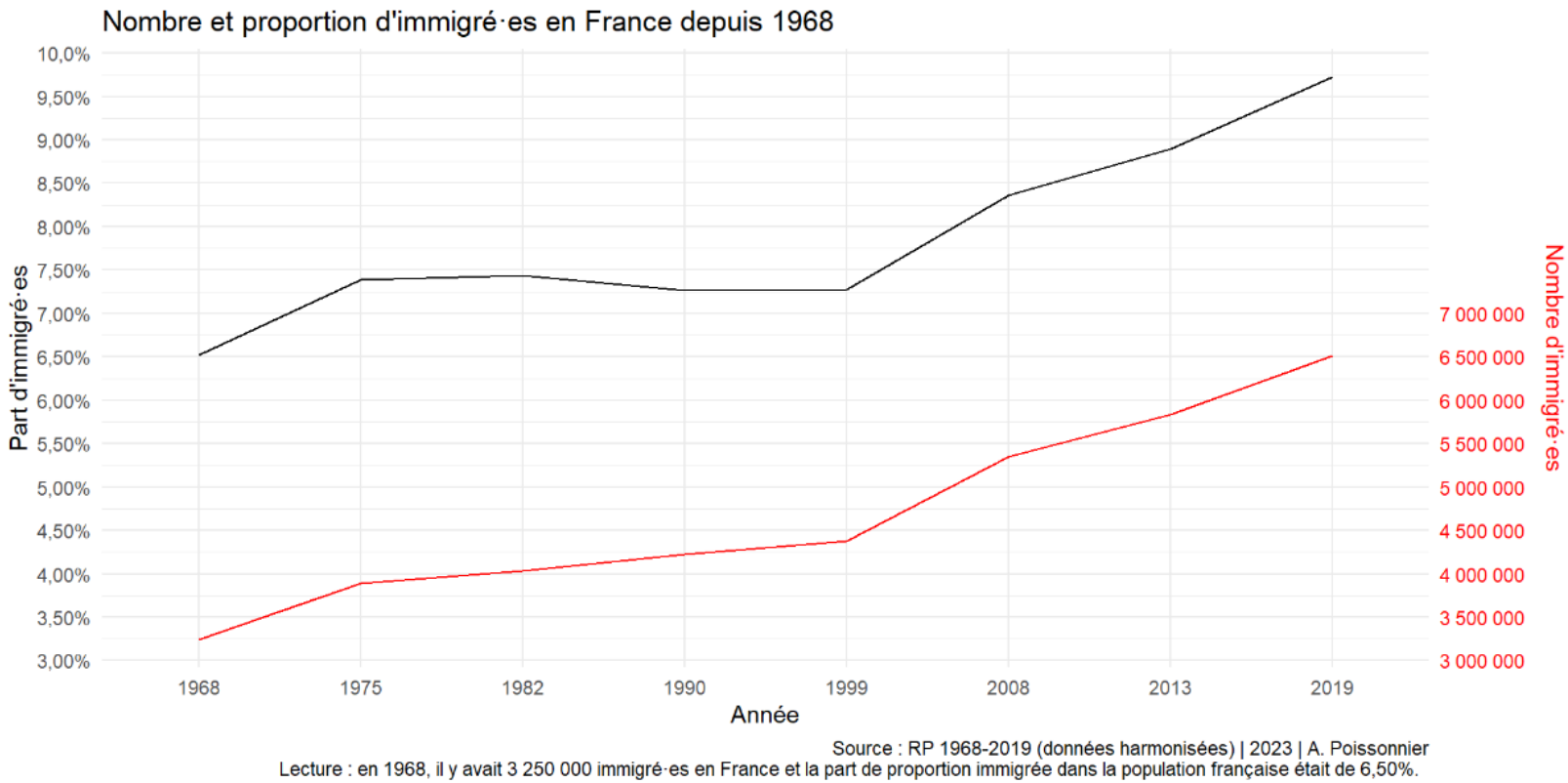
  ↳ Indiquez votre nationalité à la naissance :

- **Étrangère** 3

  ↳ Indiquez votre nationalité :

Compter : des fichiers diffusés à l'analyse statistique

En imaginant la forme du tableau de données diffusé pour le recensement, quelles opérations de collecte, de recodage et de calculs ont abouti à ce graphique ?



Rappel de la définition

Appréhender un phénomène en le décomposant en des éléments plus simples, regroupés selon des logiques d'équivalence ou de similarité. C'est une opération de quantification qui part du langage pour construire des découpages du monde et dénombrer les individus qui tombent dans ces catégories.

Le tri à plat pour compter à une dimension

Dans certains cas, le compte d'une catégorie constitue un objectif de l'analyse quantitative : connaître le nombre de personnes sans domicile fixe, mesurer la part de la population immigrée, déterminer les parts de diplômé·es du supérieur court et long.

Le *tri à plat* permet d'obtenir les *effectifs* de chaque modalité d'une variable catégorielle, c'est-à-dire le nombre d'individus présentant telle caractéristique (être diplômé·e du supérieur long) au lieu d'une ou plusieurs autres (être diplômé du supérieur court; ne pas être diplômé du supérieur).

Décomposer la population active

Le tri à plat de la population active selon la variable catégorielle du statut d'emploi distingue les personnes en emploi, les personnes dans le halo du chômage et les personnes au chômage. Il permet de calculer le taux de chômage, statistique investie d'un enjeu politique fort.

Année	Population active	Au chômage	Taux de chômage	Dans le halo autour du chômage	Taux de halo du chômage	Taux total en mal-emploi
2022	30 575	2 234	7,3	1 859	6,1	13,4
2021 (r)	30 264	2 383	7,9	1 958	6,5	14,3
2020 (r)	29 735	2 392	8,0	2 241	7,5	15,6
2019 (r)	29 963	2 527	8,4	1 890	6,3	14,7
2018 (r)	29 977	2 705	9,0	1 878	6,3	15,3
2017	29 821	2 807	9,4	1 863	6,2	15,7
r : données révisées.						
Lecture : en 2022, 2 234 000 personnes sont au chômage et 1 859 000 dans le halo autour du chômage.						
Champ : France hors Mayotte, personnes actives de 15 ans ou plus vivant en logement ordinaire.						

Figure: Taux de chômage et de halo de chômage entre 2017 et 2022

Mesurer les violences conjugales

Le tri à plat de la réponse à la question sur les violences subies au cours des 12 derniers mois dans le cadre du couple permet de mesurer la prévalence des violences conjugales. Comme ces fréquences sont calculées séparément pour les hommes et les femmes, on pourrait aussi considérer que c'est un tri croisé.

Indicateur de violence	Violence psychologique		Violence physique		Violence sexuelle		Indicateur global	
	Fem.	Hom.	Fem.	Hom.	Fem.	Hom.	Fem.	Hom.
Pas d'atteinte	95,1	98,1	96,6	99,4	99,2	99,9	93,7	97,8
Atteintes modérées	2,3	1,1	1,6	0,3	0,2	0,1	2,4	1,4
Atteintes fréquentes ou sévères	1,7	0,6	0,4	0,1	0,2	0,0	1,8	0,6
Atteintes très sévères	0,9	0,2	1,3	0,1	0,4	0,0	2,1	0,2
Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Au moins une atteinte	*4,9*	*1,9*	*3,3*	*0,6*	*0,8*	*0,1*	*6,3*	*2,2*
Effectif observé	13476	10451	13476	10451	13476	10451	13476	10451

Champ: femmes et hommes âgés de 20 à 69 ans ,vivant en France métropolitaine, en ménage ordinaire, et ayant eu au moins une relation de couple de quatre mois et plus avant les douze derniers mois, la dernière étant en cours ou terminée par une séparation ou un divorce. *Note*: les proportions de personnes ayant déclaré des atteintes (tous niveaux confondus) sont significativement différentes selon le sexe pour chaque type de violence. *Source*: enquête Virage, Ined, 2015.

Figure: Faits de violence par type et par sexe au cours de la vie conjugale avant les douze derniers mois (%)

Le tri croisé pour comparer selon deux ou plus dimensions

Dans d'autres cas, la distribution d'une variable catégorielle est peu intéressante en elle-même mais elle le devient quand on la croise avec une ou plusieurs autres variables, soit quand on compare la distribution d'une variable considérée comme *expliquée* selon une autre variable considérée comme *explicative* : la part de femmes qui déclarent que leur loisir préféré est le bricolage parmi 5 loisirs comparé à la même part chez les hommes.

Le *tri croisé* permet d'obtenir les *effectifs* pour chaque couple de modalités des deux variables catégorielles, c'est-à-dire le nombre d'individus présentant telle caractéristique (être une femme) et telle caractéristique (préférer bricoler) au lieu d'autres (être un homme et/ou préférer un autre loisir que le bricolage).

Mettre à jour des différences entre groupes

La distribution des effectifs est traditionnellement représentées en calculant des *pourcentages en ligne* ou des *pourcentages en colonne* (plus de détails à la séance 3). Les cases peuvent être comparées deux à deux ou à l'ensemble en suivant une lecture "en croix".

en %

Composition du ménage	Propriétaire	Locataire ou sous-locataire	Logé gratuitement	**Ensemble**
Personne seule	45,4	51,8	2,8	**100,0**
Couple sans enfant	74,6	24,0	1,4	**100,0**
Couple avec enfants	67,8	30,4	1,8	**100,0**
Famille monoparentale	37,9	60,2	2,0	**100,0**
Autre type de ménage	31,8	64,4	3,7	**100,0**
Ensemble	**57,6**	**40,2**	**2,1**	**100,0**

Lecture : Au 1er janvier 2020, 45,4 % des ménages constitués d'une personne seule sont propriétaires de leur résidence principale.
Champ : France métropolitaine.
Source : Insee, recensement de la population 2020 (exploitation complémentaire).

Figure: Répartition des ménages par statut d'occupation selon la composition du ménage au 1er janvier 2020

Ici les pourcentages en ligne permettent de connaître la distribution du statut d'occupation du logement pour chaque sein de type de ménage.

Commencer par la lecture

La lecture doit toujours commencer par citer les chiffres, les comparer puis conclure.

- En 2020, 51,8% des personnes seules étaient locataires ou sous-locataires contre 30,4% pour les couples sans enfants. Les personnes seules sont donc **plus souvent locataires** que les couples sans enfants.
- En 2020, 45,4% des personnes seules étaient propriétaires de leur logement contre 57,6% dans l'ensemble. Les personnes seules sont donc **sous-représentées** parmi les propriétaires.
- En 2020, 67,8% des couples avec enfants étaient propriétaires de leur logement contre 57,6% dans l'ensemble. Les couples avec enfants sont donc **sur-représentés** parmi les propriétaires.

La première partie de la lecture implique donc d'utiliser l'indicatif et d'être précis·e sur les informations données.

Puis interpréter

La lecture est ensuite suivie d'une interprétation, fondée sur votre intuition sociologie ($\neq$ opinion personnelle), sur d'autres statistiques ou encore sur d'autres travaux.

- Le fait que les personnes seules sont (indicatif) plus souvent locataires que les couples pourrait (conditionnel) s'expliquer par le fait que les projets d'achat soient (subjonctif) souvent conçus au moment où la famille se forme.
- Les conditions d'accès au crédit (Gollac, 2011) et le modèle familial de la propriété immobilière (Bonvalet, 2007) peuvent expliquer (indicatif) que les personnes seules sont (indicatif) plus souvent locataires que les couples. En 2017, 75% des primo-accessions étaient réalisées par des couples avec ou sans enfants (INSEE, 2017). La majorité des ménages qui deviennent propriétaires sont donc des ménages conjugaux.

Étudier la composition sociale d'un sous-groupe

Les pourcentages en colonne permettent ici de connaître la composition sociale des Grandes Écoles.

Origine sociale des élèves des grandes écoles en France 2016-2017 (en %)

Distribution des élèves des grandes écoles par profession et catégorie socio-professionnelle (PCS) de leurs parents en France pour l'année

Catégorie sociale	Écoles d'ingénieur	Écoles de commerce	École normale supérieur	Instituts d'études politiques	Autres grandes écoles
PCS défavorisées	9	7	6	8	9
PCS moyennes	17	18	14	14	18
PCS favorisées	11	10	8	10	10
PCS très favorisées	63	65	72	68	63
Ensemble	100	100	100	100	100

Source : Minètre de l'Enseignement supérieur et de la Recherche, 2021, Statistica

Champ : élèves inscrits dans une formation "Grandes écoles" en 2016-2017

Lecture : en 2016, 9% des étudiant·es des écoles d'ingénieur étaient issu·es des PCS défavorisées.

- En 2016, 6% des étudiant·es des Écoles Normales Supérieures étaient issues des PCS défavorisées contre 72% pour les PCS très favorisées. La majorité des étudiant·es des ENS sont donc issu·es des classes supérieures hautes.
- On ne connaît pas ici la part des étudiant·es du supérieur issu·es des PCS défavorisées inscrit·es dans une ENS !

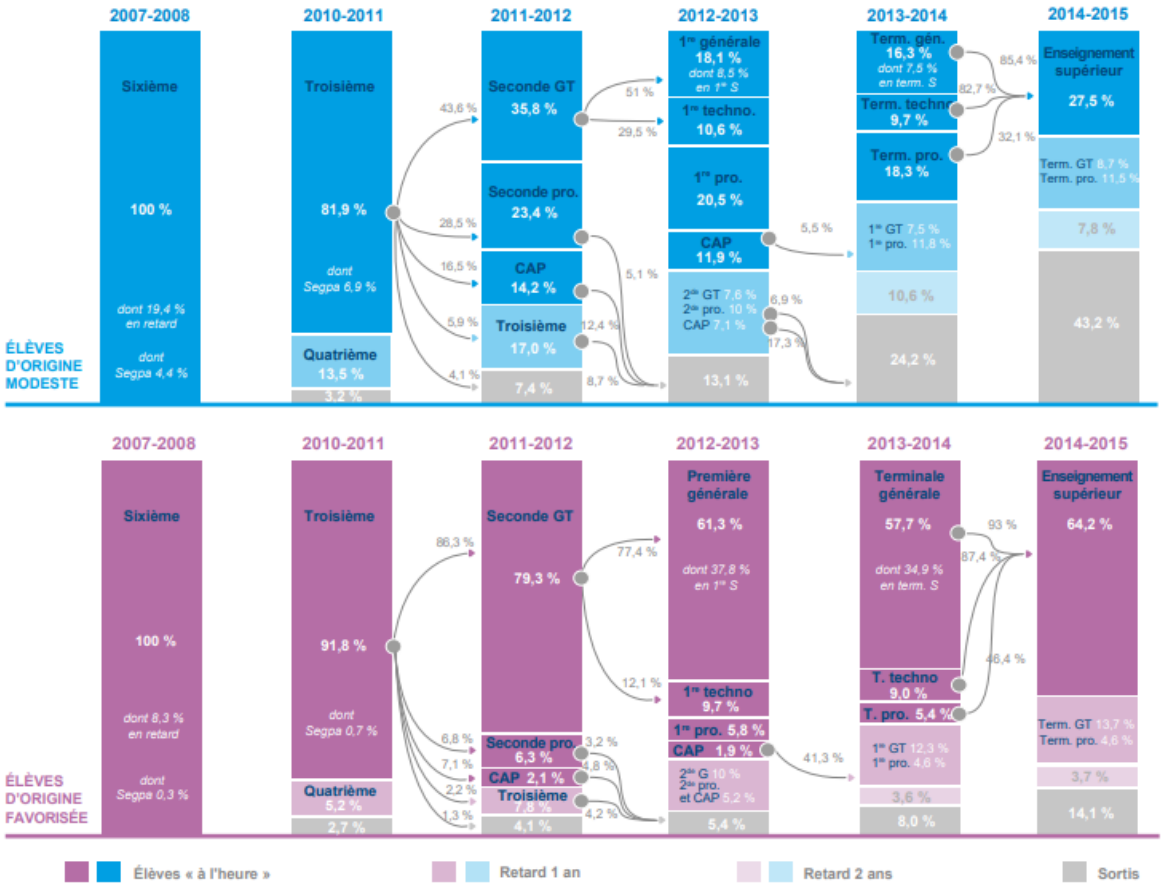
En ligne ou en colonne ?

Une vérification pour ne pas se tromper

- Il faut toujours chercher où est-ce que la somme des pourcentages fait 100%, soit en cherchant la case Ensemble ou Total, soit en faisant soi-même la somme.
- Les variables explicatives ne sont pas toujours mises en lignes et celles expliquées en colonnes : le sens d'un calcul des pourcentages en ligne ou en colonne change donc selon la disposition du tableau croisé.
- D'où l'importance de la note de lecture, sur laquelle nous reviendrons dans la prochaine séance.

D'un tableau croisé à l'autre

Graphique 2 – Trajectoires comparées de la sixième à l'enseignement supérieur, selon l'origine sociale



Notes :
- sont prises en compte ici les PCS ménage (nomenclature des professions et catégories socioprofessionnelles des ménages) ;
- sont considérés d'origine modeste (35 % des élèves) les enfants de ménages à dominante ouvrière ou composés d'une personne ouvrière ou employée sans conjoint ou avec un conjoint inactif. Sont considérés d'origine favorisée (30 % des élèves) les enfants de ménages à dominante cadre ou intermédiaire/cadre ;
- en sixième, sont dits en retard les élèves ayant redoublé au moins une année en primaire ;
- pour les autres années, sont dits en retard d'un ou de deux ans les élèves ayant redoublé ou ayant été réorientés dans l'enseignement secondaire ;
- Segpa : section d'enseignement général et professionnel adapté. Seconde GT : seconde générale et technologique.

Lecture : en 2010-2011, 81,9 % des enfants de famille modeste entrés en sixième en 2007 sont en troisième, 13,5 % sont encore en quatrième et 3,2 % sont sortis du système éducatif. Le faible écart entre cette somme et 100 % s'explique par l'existence d'élèves « perdus » dans le panel, décédés ou en très grand retard, ayant sauté une classe ou suivant une autre scolarité. Les flèches indiquent les taux de passage : 86,3 % des élèves de famille favorisée et 43,6 % des enfants de famille modeste qui étaient en troisième en 2010-2011 sont allés en seconde générale et technologique en 2011-2012. Pour simplifier, seuls quelques taux significatifs sont représentés.

Source : calculs France Stratégie à partir des données du panel 2007 (DEPP)

Des notions proches mais différentes

- La démarche sociologique vise traditionnellement à expliquer un fait social à partir de déterminants sociaux (approche macro d'inspiration structuraliste) ou du sens données par les individus à leurs actions sociales (approche micro liée à la démarche compréhensive).
- Le lien entre deux variables qualitatives est normalement appelé *dépendance*.
- Le lien entre deux variables qualitative est appelé *corrélation*.
- La mesure empirique des dépendances ou corrélations est un préalable à l'identification de liens de *causalité*, c'est-à-dire la situation où une variable X modifie une variable Y, mais elle ne suffit pas.

Une fausse relation de causalité ...

Une enquête sur les accidents de la route est organisée en France. Les données permettent de conclure que les parisien·nes ont moins souvent eu d'accidents de voiture que les habitant·es du reste du territoire. Peut-on conclure que les parisien·nes sont de meilleur·es conducteur·ices que les autres habitant·es du territoire ?

Une fausse relation de causalité ...

Une enquête sur les accidents de la route est organisée en France. Les données permettent de conclure que les parisien·nes ont moins souvent eu d'accidents de voiture que les habitant·es du reste du territoire. Peut-on conclure que les parisien·nes sont de meilleur·es conducteur·ices que les autres habitant·es du territoire ?

liée au biais de variable omise

Les parisien·nes ont sûrement moins d'accidents de voiture car ils et elles prennent bien moins souvent la voiture que les transports en commun. La dépendance entre la variable géographique (X) et la variable d'accidents (Y) est bien empirique vérifiée, mais c'est le nombre de kilomètres parcourus en voiture (X') qui doit expliquer la majorité de ce lien.