Analyse statistique

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2023

Statistiques et sciences sociales

Les méthodes qualitatives et quantitatives visent toutes deux à produire des matériaux empiriques mobilisés dans une démarche scientifique. Les techniques de quantification du monde social sont le fruit de plusieurs histoires dont celles :

- Des États modernes.
  A. Desrosières, 1993, *La politique des grands nombres. Histoire de la raison statistique*
- Des disciplines scientifiques et de leurs paradigmes.
  O. Martin, 2002, « Mathématiques et sciences sociales au XXème siècle » dans *Revue d'histoire des sciences humaines*
- De la réception et des usages sociaux des statistiques.
  I. Bruno, E. Didier et J. Prévieux, 2014, *Statactivisme*

La statistique d'État, les statistiques de l'État

La statistique s'est institutionnalisée comme une science de gouvernement : au XIXe, la construction des États modernes accompagne le développement d'un appareil statistique et l'invention d'instruments de compte, de description de la société et de prévision.

- Compter la population et définir la citoyenneté grâce au recensement
- Identifier les facteurs de fécondité et mener des politiques natalistes à partir de la démographie
- Développer l'État-providence et ses dispositifs de protection collective en menant des enquêtes sur les conditions de vie et de travail
- Soutenir la croissance économique par la mesure de l'activité

L'usage scientifique des statistiques

Les disciplines ont développé des méthodes statistiques qui se sont cristallisées dans la pratique scientifique. L'évolution du paradigme scientifique passe aussi par l'imposition de nouveaux outils de quantification, ou la disqualification d'outils auparavant dominants.

- La comparaison de moyennes avec un groupe de contrôle et un groupe traité est devenue la norme dans les sciences médicales
- L'économétrie et les techniques statistiques élaborées dominent les sciences économiques
- Les sociologues font des statistiques descriptives sur leurs populations et recourent aux grandes enquêtes pour identifier des régularités sociales, et ce selon des traditions nationales : aux États-Unis, les méthodes quantitatives économétriques sont bien plus dominantes qu'en France où l'usage des statistiques est plus diffus et sélectif

La réception et les usages sociaux des statistiques

Les statistiques sont aussi reçues, mobilisées et produites par les agents sociaux pour émettre des énoncés descriptifs et souvent critiques à partir de "faits".

- Mesure des discriminations et des inégalités, portée par des groupes militants et scientifiques
- Étude des conditions de vie et des risques spécifiques à certaines classes
- Remise en cause des indicateurs ou mise en perspective de leur interprétation

Vérité statistique ou appauvrissement de la réalité

L'opération de *quantification* pourrait être définie comme la construction d'une représentation du monde social sous forme de chiffres, en particulier quand les dimensions représentées ne sont pas naturellement numériques.
De manière schématique, la vision de la quantification comme un moyen d'accès à des vérités statistiques s'oppose à celle qui affirme que les statistiques sont nécessairement partielles et orientées, voire trop facilement manipulables.

- L'exemple classique de la critique de l'analyse quantitative est celle portant sur les sondages
- La socio-histoire de la quantification fait la généalogie des techniques aujourd'hui dominantes, révélant les intérêts et les contextes qui ont permis leur imposition
- Même les statistiques simples et familières posent des difficultés qui deviennent apparentes quand on se confronte aux étapes de leur production

Les types de variables

La qualité - nominales, catégorielles ou qualitatives

Nommer ou de dénommer une « qualité », c'est-à-dire d'associer une étiquette, une catégorie à un individu. Pour désigner les catégories d'une variable nominale, on parle de modalités de variable.

L'ordre - ordinales, discrètes ou d'échelle

Variables catégorielles spécifiques dont on peut hiérarchiser les catégories : les classes d'âge, découpées par intervalle de dix ans; les échelles de satisfication; les mentions du baccalauréat.

La quantité - numériques ou quantitatives

Les variables dont la mesure a du sens, celles-ci ayant toujours une unité de mesure : revenu moyen des ménages en euros; la taille moyenne des logements en mètres carrés. L'âge, lorsqu'il est exprimé en continu (1 an, 2 ans, 3 ans, etc.) et non en tranche, est aussi une variable quantitative.

Catégories et indicateurs

Les *catégories* et les *indicateurs* sont deux constructions statistiques à la base de l'analyse quantitative.

La catégorisation

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.

Les indicateurs

Décrire les caractéristiques d'un individu ou d'une population à partir d'une valeur numérique synthétique. C'est une opération de quantification qui transforme la mesure d'une quantité pour rendre compte de sa distribution, cette dernière disant quelque chose de l'organisation du monde social.

La moyenne arithmétique

La *moyenne* $\overline{X}$ est égale à la somme des valeurs X_i prises par une variable divisée par le nombre N d'individus statistiques (= une personne, mais aussi un ménage, une école, un pays)

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{1}$$

La moyenne arithmétique pondérée

On peut vouloir pondérer la moyenne pour donner une importance différente à chaque individu. Par exemple, la moyenne d'un bulletin scolaire comporte souvent des coefficients qui font que les mathématiques comptent plus que le cours d'éducation musicale. La *moyenne pondérée* $\overline{X_p}$ est égale à la somme des valeurs X_i prises par une variable, multipliées par le poids P_i, divisée par la somme des poids.

$$\overline{X_p} = \frac{\sum_{i=1}^{N} X_i \times P_i}{\sum_{i=1}^{N} P_i} \tag{2}$$

Les quantiles

Les *quantiles* correspondent aux valeurs particulières d'une variable qui divisent la population en intervalles d'une amplitude donnée.

- Les *déciles* découpent la population en 10 parts égales
  Les 10% les plus riches (revenus $>$ 9ème décile) contre les 10% les plus pauvres (revenus $<$ 1er décile)
- Les *quartiles* découpent la population en 4 parts égales
  Les 25% les plus riches (revenus $>$ 3ème quartile) contre les 25% les plus pauvres (revenus $<$ 1er quartile)
- La *médiane* découpe la population en 2 parts égales
  Les 50% les plus riches (revenus $>$ médiane) contre les 50% les plus pauvres (revenus $<$ médiane)

Le mode

Le *mode* correspond à la valeur la plus typique de l'ensemble, c'est-à-dire celle qui apparaît le plus souvent.

Application : l'âge au décès

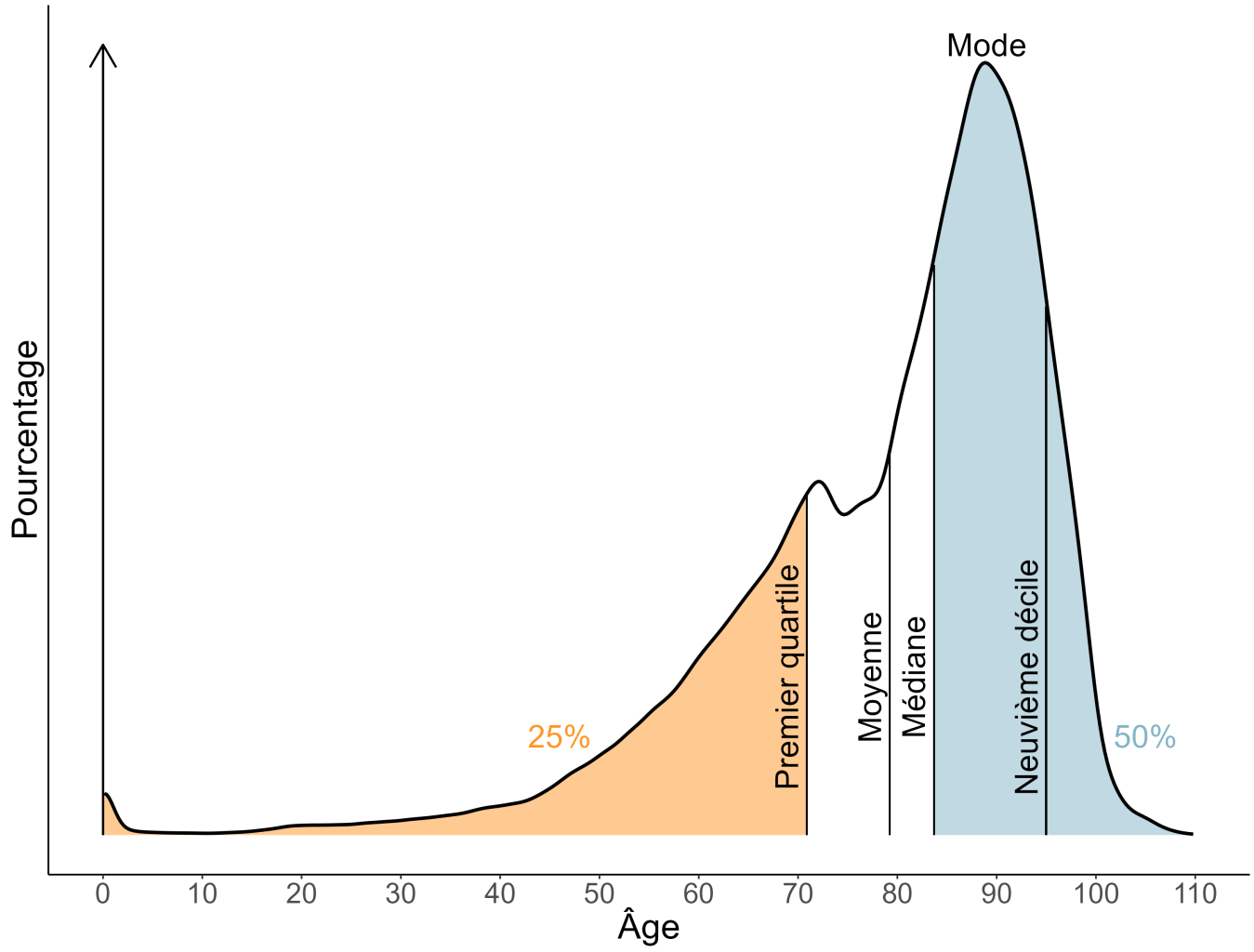


Figure: Distribution de l'âge des personnes décédées en 2019 en France

L'étendue

L'*étendue* est la différence entre la plus petite valeur et la plus grande.

L'écart interquartile

L'*écart interquartile* est l'étendue entre le premier et le troisième quartiles, soit les 50% des données qui sont au centre de la distribution.

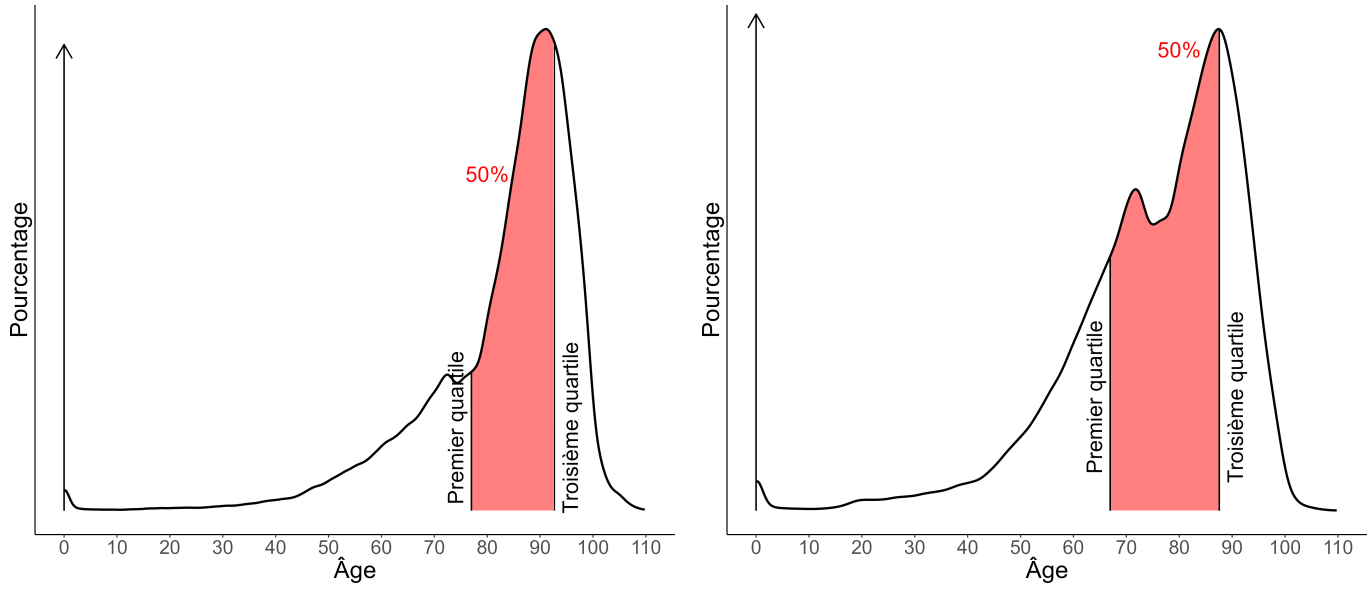
La variance

La *variance* σ est la moyenne au carré des écarts entre chaque valeur X_i prise par une variable et la moyenne $\overline{X}$.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N} \tag{3}$$

L'*écart-type* σ est la racine carrée de la variance.

Application : l'âge au décès



(a) Femmes (b) Hommes

Figure: Distribution de l'âge des personnes décédées en 2019 en France selon le sexe

Application : l'espérance de vie

Comment est-ce que vous définiriez l'espérance de vie et, selon vous, quel est le lien entre son calcul et la distribution des âges au décès ?

Définition

L'espérance de vie, c'est l'âge auquel vivrait une personne si les conditions de mortalité observées cette année-là se maintenaient tout au long de sa vie.
Le calcul de l'espérance de vie repose sur deux opérations : le calcul d'un taux de mortalité pour chaque âge, la modélisation d'une génération fictive. L'espérance de vie correspond à l'âge moyen au décès de cette génération fictive : on applique le taux de mortalité âge par âge jusqu'à ce que toute notre population soit décédée, et on fait la moyenne arithmétique de l'âge de décès.

Application : l'espérance de vie

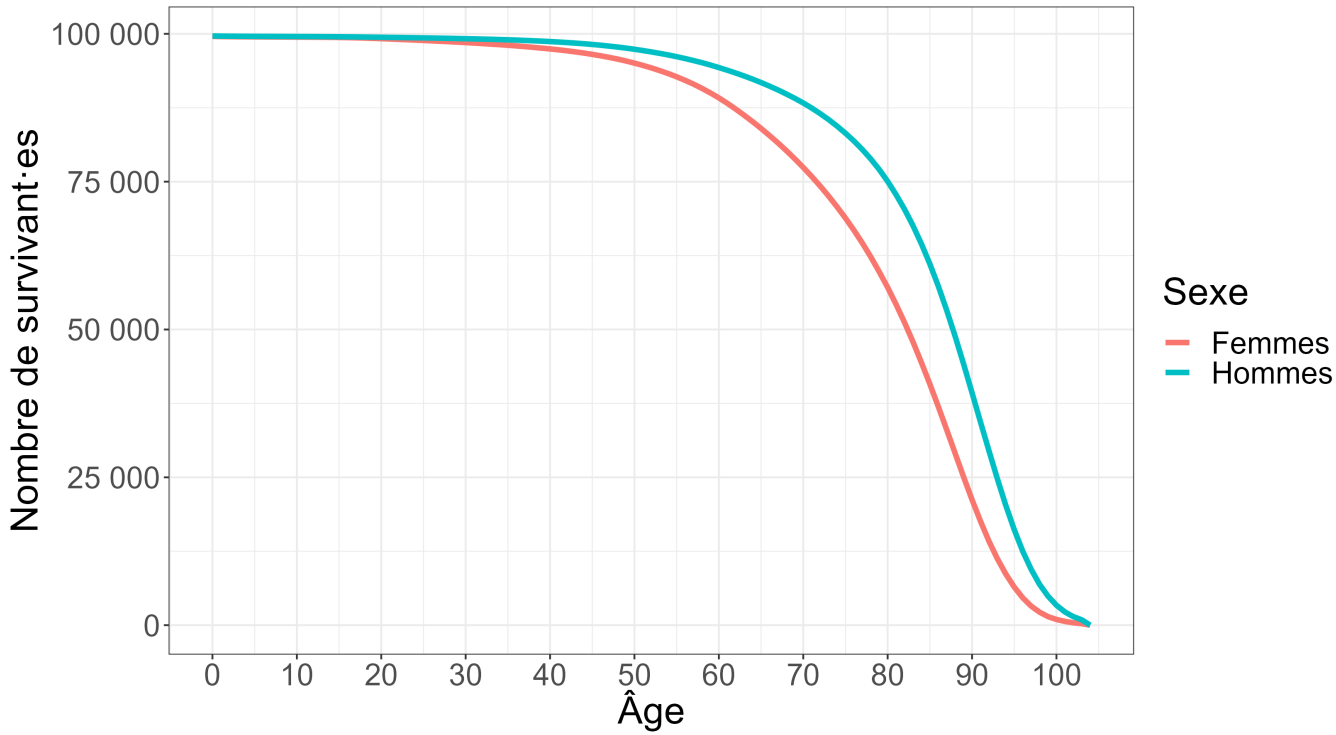


Figure: Nombre de survivant·es selon l'âge pour une génération fictive née en 2019

Pourquoi une génération fictive ?

Pourquoi ne prend-on pas simplement la moyenne des âges des personnes décédées dans la réalité pour affirmer, par exemple, qu' « en moyenne, en France, on meurt à 79 ans » ?
La valeur de cet indicateur dépend de la structure des âges de la population. Une population âgée tendra à avoir un âge moyen de décès supérieur à une population avec une part importante de jeunes adultes. L'âge moyen au décès ne permet donc pas de modéliser la trajectoire attendue d'une personne née une année donnée.

Un indicateur se doit d'être synthétique

L'indicateur d'espérance de vie permet d'utiliser un seul chiffre pour résumer plus d'une centaine de taux de mortalité par âge mesurés pour une année donnée, sans attendre de connaître l'âge réel de décès de la génération étudiée. Il résume donc un ensemble d'informations complexes, qui peuvent être interprétées grâce à cette statistique unique, mais implique nécessairement une simplification de la réalité.

L'interprétation des indicateurs

C'est en général la comparaison qui permet d'interpréter un indicateur

- Les écarts d'un indicateur entre plusieurs catégories
- L'évolution dans le temps d'un indicateur pour une même population
- La comparaison entre plusieurs indicateurs

On tente alors de rendre compte de la *régularité* et de la *variabilité*

- Régularité : sociologie durkheimienne, les statistiques permettent de passer de l'individuel au collectif en observant les caractéristiques du groupe, ce qui donne accès au social à défaut de pouvoir expérimenter
- Variabilité : approfondi par Maurice Halbwachs, qui s'oppose à la théorie de l'homme moyen d'Aldophe Quetelet et s'intéresse aux écarts entre catégories et aux évolutions des statistiques plutôt qu'à leur synthèse

Application : l'espérance de vie

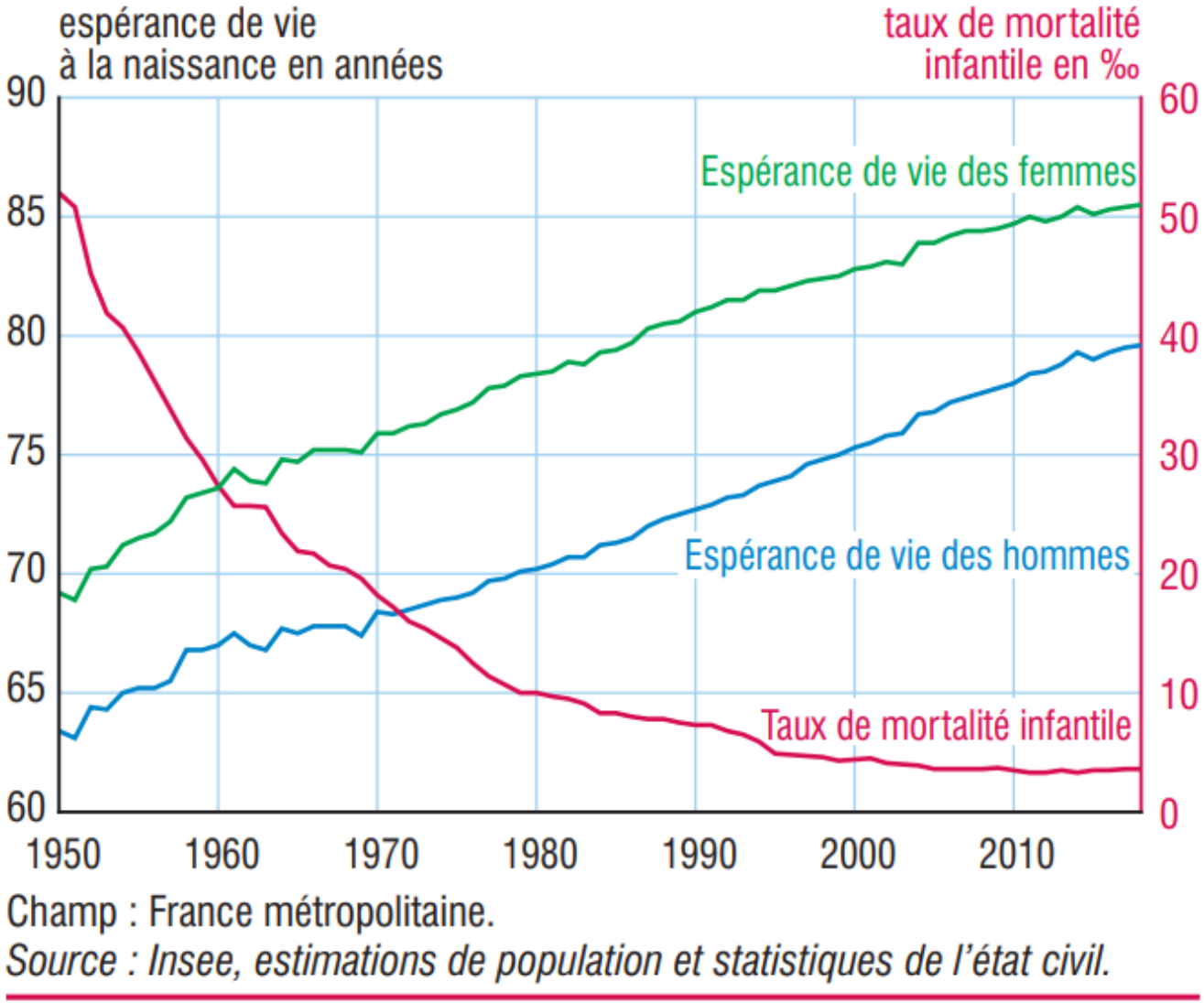


Figure: Espérance de vie à la naissance et taux de mortalité infantile de 1950 à 2019

Application : l'espérance de vie

Montant total des retraites perçues par catégorie sociale pour les hommes

	Âge moyen de la retraite* (A) en ans	Espérance de vie** (B) en années	Espérance d'années de retraite (B-A) en années	Montant des pensions de retraite perçues au cours de la vie*** en euros
Ouvriers	61,8	77,6	15,8	285 680
Employés	62,7	79,9	17,2	310 684
Professions intermédiaires	61,8	81,7	19,9	518 530
Cadres supérieurs	63,1	84,0	20,9	930 117

*Source : IPP, données 2018-2020. **Mesurée à 35 ans. Source : Insee, données 2009-2013. ***Estimation Observatoire des inégalités
Lecture : au cours de sa vie, un employé peut espérer toucher 310 684 euros de pensions de retraite en moyenne, selon les estimations de l'Observatoire des inégalités.
Source : estimation de l'Observatoire des inégalités, d'après les données de l'IPP et de l'Insee – © Observatoire des inégalités

Figure: Conditions de sortie de la vie active selon la catégorie socio-professionnelle